MASARYK UNIVERSITY

FACULTY OF SCIENCE

RECETOX

Integrative Bioinformatics and Computational Modelling in Colorectal Cancer: Unveiling Tumor Heterogeneity through Multi-Omics Data

HABILITATION THESIS

MGR. EVA BUDINSKÁ, PH.D.

BRNO 2025

ACKNOWLEDGEMENTS

First and foremost, I want to thank my best friend, love of my life, and husband, Vlad. None of this would have been possible without his unwavering support, his willingness to share in everything—from raising our kids to managing our home—and his endless patience with me. I am beyond grateful to have him by my side.

A huge thank you also goes to my family—my mom, dad, and mother-in-law—whose help with our children has been absolutely invaluable. Without them, juggling science and family life would have been so much harder.

And to my children—Anna, Mihai, and Adrian—you bring so much joy, laughter, and meaning to my life. Anna and Mihai, you have been my greatest motivation, even if you sometimes make finishing a sentence (or a thesis) a challenge. And to Adrian, though you are already carving your own path in life, I am grateful for the years I have known you and for the moments we have shared as a family. You were always so wonderfully curious as a child, and it has been a privilege to watch you grow into the person you are today.

I am also deeply grateful to my colleagues and collaborators, who have made this journey not only productive but also enjoyable. Their insights, dedication, and teamwork have been instrumental in shaping this work, and I truly appreciate the many discussions, shared frustrations, and moments of celebration along the way.

Last but by no means least, I want to express my deep gratitude to Prof. Jana Klánová for giving me the opportunity to do my research in such a supportive and inspiring environment. Her guidance and encouragement have meant the world to me, especially in the moments when I needed it most.

FUNDING

This work received funding from the Ministry of Health of the Czech Republic, grant nr. 16-31966A and 19-03-00298. All rights reserved.

The author acknowledges funding from the Czech Science Foundation (GAČR) grant no. GA19-08646S

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825410 (ONCOBIOME project). This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

The work was supported by RECETOX research infrastructure through project LM2018121 funded by the Czech Ministry of Education, Youth and Sports, and the support of CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 (857560) and the Czech Ministry of Education, Youth and Sports (02.1.01/0.0/0.0/18_046/0015975)

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

CONTENT

1. OVERVIEW

Colorectal cancer (CRC) is a highly heterogeneous disease, both at the molecular and cellular levels. This heterogeneity significantly influences tumor progression, therapy responses, and clinical outcomes. Fortunately, the advances in multi-omics technologies and computational methods provided new opportunities for comprehensive investigation. However, multi-omics data generated from high-throughput technologies are particularly sensitive to technical variability and batch effects. Producing meaningful and robust results requires careful preprocessing of raw data and highly specialized expertise in data mining. This thesis leverages my unique expertise in bioinformatics and computational modeling to analyze and integrate multi-omics data, with the aim of uncovering tumor-specific molecular patterns and the complex interactions within the tumor microenvironment.

The thesis is organized into thematic areas reflecting the scope of my research contributions. While foundational work in **computational methodology** (chapter 3.1) and **preclinical models** (chapter 3.3) is included mostly for context, the primary emphasis is placed on **molecular subtyping and tumor heterogeneity** (chapter 3.2), the **integration of imaging and omics data** (part 4), the **tumor microenvironment and microbiome** (chapter 3.5), and **clinical applications in diagnostics and therapy** (chapter 3.6). These themes are supported by results from key publications that demonstrate how computational approaches in multiomics setting provide novel insights into CRC biology.

The development and application of **computational tools** form the foundation of my research. This section highlights efforts to create robust bioinformatics pipelines and tools that enable effective multi-omics data analysis and interpretation. For instance, in **[*1*]** we introduced `Rgtsp`, a generalized top-scoring pairs package that enabled class prediction in gene expression datasets, setting a foundation for subsequent predictive modeling. Expanding on these efforts, in [*2*] we presented `TopKLists`, an R package designed for statistical inference and aggregation of ranked omics datasets, addressing challenges in integrating heterogeneous high-dimensional data. Similarly, in our work **[*3*]** we introduced **ToPASeq**, a novel package that implements six methods for topological analysis of RNA-Seq and microarray data analysis. Finally, in **[*4*],** we leveraged this R package and critically compared topology-based pathway analysis methods, evaluating their consistency and biological inference across diverse datasets. Collectively, these tools not only provide a methodological basis for subsequent studies but also serve as valuable resources for the broader scientific community, facilitating reproducibility and innovation in data-driven cancer research.

The chapter on **molecular subtyping and tumor heterogeneity** focuses on the identification and characterization of molecular subtypes in CRC, including their clinical relevance. This includes studies such as [7] and [9] in which we identified gene expression-based CRC subtypes and linked them to prognosis and treatment responses. In [5], we provided a comprehensive characterization of genome-wide copy number aberrations in CRC, revealing novel oncogenes and distinctive alteration patterns relevant to tumor heterogeneity. Furthermore, in [6], we examined differences between distal and proximal colon cancers, uncovering molecular, pathological, and clinical features that distinguish these CRC subtypes. In [8], we assessed the prognostic role of BRAF and KRAS mutation in the context of the tumour sidedness and MSI status. Last, we investigated tumor architecture and morphological heterogeneity, providing insights into how structural and molecular variations within tumors affect clinical outcomes [1]. These results collectively highlight the importance of understanding CRC at the molecular and structural levels to refine therapeutic strategies.

Preclinical models are instrumental in bridging the gap between computational insights and biological validation. This theme focuses on cross-species analyses and experimental systems that enhance our understanding of cancer biology. In [10], we investigated molecular hallmarks of colorectal cancer using genetically engineered mouse models, identifying parallels with human disease at the transcriptomic level. Complementing this work, we emphasized the utility of patient-derived xenografts (PDX) in precision oncology, illustrating how computationally derived hypotheses can be tested in biologically relevant systems [11]. Additionally, preclinical efforts in projects like these have enriched the understanding of tumor evolution and therapeutic response, highlighting the synergy between computational modeling and biological experimentation.

The **integration of digital pathology and omics data** represents a critical advancement in CRC research, as it bridges spatial and molecular heterogeneity. In [12], we contributed to this field by showing how joint analysis of histopathology images and transcriptomic data can yield biomarkers for molecular subtypes in breast cancer. Consequently, in [13] we were the first to demonstrate how histopathological image features combined with gene expression data enable deeper insights into CRC biology. Building on this, in [14] we examined gene expression signatures within macro-dissected spatially resolved tumor regions, uncovering specific spatially distributed molecular patterns. These integrative approaches have proven effective in identifying image-based molecular biomarkers, advancing precision oncology.

The role of the **tumor microenvironment and microbiome** in CRC progression is presented in chapter 3.5. Our work, published in **[*16*],** demonstrated how stool sampling techniques influence microbiome composition, emphasizing the importance of methodological consistency in microbiome studies. This knowledge was further applied in the analysis of microbiome data of the Colobiome (AZV) study, which resulted in identification of distinct microbiome-defined CRC subtypes that correlate with tumor characteristics, revealing how microbial signatures associate with tumor progression **[*17*]**. In **[*18*],** we describe how microbial signatures correlate with tumor progression and immune modulation. These findings show the importance of environmental and microbial factors in CRC biology and their potential for biomarker development.

Finally, the thesis emphasizes the practical applications in **clinical diagnostics and treatment selection**. In **[*19*]**, we explored mRNA biomarkers for assessing FOLFIRI treatment efficacy in Stage III colon cancer, demonstrating the potential for optimizing therapy selection based on molecular profiling. Next, we derived fecal microRNA signature for CRC diagnosis [***20***] and a gene expression signature for identifying high-risk stage IIA CRC patients mining molecular data from macrodissected invasion front area **[*21*]**. These studies show the translational potential of computational approaches to improve patient care.

By presenting results across these interconnected themes, this thesis provides a cohesive overview of CRC heterogeneity and demonstrates the role of integrative computational methods in advancing both basic and clinical cancer research in CRC.

Technical note: Throughout the text, references are cited using numbered brackets **[ ]**, with each number corresponding to the full citation in the reference list at the end of the thesis, presented in the order as they appear in the text. To distinguish between references to my own publications and those from other sources, references to my publications are numbered according to the list of my works provided in this thesis and are ***emphasized*** within the text. For example, a reference to the first entry in the list of my referenced publications appears as [*1*], whereas a reference to other sources is cited as [1]. This system ensures clarity when referring to my contributions in the referred list of publications in comparison to my other contributions (links to SW, patents, preprints of articles) or external works.

2. INTRODUCTION

Modern cancer research is inherently multidisciplinary, employing multi-omics approaches to study cancer heterogeneity from different perspectives and is thus heavily reliant on computational, statistical, and bioinformatics approaches. Among solid cancers, colorectal carcinoma (CRC) is one of the most common, representing a significant global health burden. Globally, CRC is the third most diagnosed cancer, accounting for 9.6% of all new cancer cases, following lung and breast cancers. It is also the second leading cause of cancer-related mortality, responsible for 9.3% of all cancer deaths, with over 1.9 million new cases and 904,000 deaths reported in 2022 [2]. In Europe, CRC is the second most frequently diagnosed cancer, making up approximately 12% of all cancer cases, with over 500,000 new cases diagnosed annually and around 250,000 deaths each year [2]. In the Czech Republic, CRC is particularly prevalent, consistently ranking among the top cancers in both incidence and mortality. Recent data indicate that the age-standardized incidence rate for CRC in Czech men is among the highest globally, ranking 13th worldwide and 12th in Europe, while for Czech women, it ranks 21st worldwide and 14th in Europe. The mortality-to-incidence ratio (M/I) for CRC in the Czech Republic is approximately 0.42, reflecting ongoing challenges in early detection and treatment [3].

Most importantly, CRC is among the most heterogeneous solid cancers, exhibiting extensive variability at the molecular, cellular, histopathological and clinical levels, including response to therapy [4–6]. The current standard treatments remain ineffective for a large group of CRC patients due to inappropriate patient selection. This means the patients are subjected to unneeded toxic treatments and that overall costs are too high with respect to achieved efficiency. Therefore, the identification of predictive biomarkers of clinical response is an absolute requirement for personalizing the treatment, with numerous benefits for the patients and the health care system. This translates into an urgent need for robust disease subclassifiers, that can explain the clinical heterogeneity of CRC beyond the currently used clinical risk factors (bowel obstruction and perforation, T4 tumour, presence of lymphovascular or perineural invasion, …) and molecular markers (such as microsatellite instability - MSI, or mutations in known oncogenes - *KRAS* or *BRAF)*. The state-of-the-art approach to bridge this gap is tumour molecular profiling. Indeed, evidence of clinically relevant tumour molecular heterogeneity has been flowing from high-throughput gene expression and mutation analyses, copy number variation assessment, methylation, miRNA and proteomic studies [7,8] [**5**]. The molecular profiling research has taken two main routes: The **supervised** approach stems from comparison of known groups (i.e. patients with early vs. late relapse of the disease) and aims at either explaining differences between groups by identification of the „affected“ molecular pathways, or searches for surrogate and measurable predictive or prognostic *signatures*, that would serve as decision tools in personalized medicine. The **unsupervised** approach, on the other hand, recognizes the molecular phenotype as an additional piece of the puzzle that **complements the complex picture of tumour heterogeneity**, which may or may not be correlated with known clinical risk factors, prognosis or response to therapy. Several studies deriving **unsupervised CRC gene expression subtypes** and stratifying tumour aggressiveness and response to treatment were published and led to the definition of four consensus molecular CRC subtypes [9].

This heterogeneity underscores the need for multidisciplinary approaches and advanced methodologies to unravel its complexities and improve patient outcomes [4].

3. MAIN TEXT

3.1. COMPUTATIONAL METHODOLOGY

The rapid growth of multi-omics technologies has necessitated the development of computational methods capable of handling large, high-dimensional datasets. Since then, computational methodologies play a crucial role in uncovering hidden patterns and relationships within complex biological data and the integration of data across molecular, spatial, and temporal domains requires innovative algorithms to ensure meaningful biological interpretations. As a mathematical biologist by training, my research focuses on using these skills to advance our understanding of cancer biology. While the development of computational tools is an integral aspect of my expertise, my primary motivation lies in applying them to translational cancer research. I advance methodologies and develop bioinformatics, data mining, and image analysis tools, when necessary, driven by the biological and clinical questions of my research, which aims at uncovering novel insights into colorectal cancer heterogeneity and its implications for diagnosis, treatment, and patient outcomes.

Class prediction

Development of clinically applicable biomarkers is usually a key focus of clinically oriented cancer research and requires identification of molecular or multi-omics signatures that are transferable across platforms and suitable for integration into routine clinical practice. Interpretability is a critical aspect of computational tools, particularly for applications in clinical decision-making. Achieving this often requires the use of explainable classification approaches that rely on a limited number of features.

One of our early works [***I***] focused on the development of methodology for explainable class prediction, applicable beyond gene expression datasets. In this study, we developed a computational tool designed to enhance class prediction across various datasets, including gene expression profiles. This methodology centers on the Top Scoring Pairs (TSP) classifier [10], which utilizes relative ranking of variable pairs to predict class labels. By focusing on the relative expression ordering of gene pairs, the method offers robustness against technical variations across different platforms, making it particularly suitable for developing clinically applicable biomarkers. Our contributions include a parallel implementation of the TSP classifier to significantly reduce training time and extensions to handle multi-class classification problems. The Rgtsp package, implemented in C++ with R functions, offers functions for k-fold cross-validation and proposes using classification trees built on top of TSP predictions for multi-class problems. This methodology has been implemented as an R package [11], which is freely accessible to the research community in GitHub [12]. This classifier was subsequently employed in [7], where we identified and characterized a subgroup of colorectal cancers that shared molecular and clinical features with BRAF-mutated tumors, despite not harboring the actual BRAF mutation. This involved developing an explainable classification system to stratify tumors into a BRAF-like subtype, as further detailed in Part 2 of this thesis**.**

Ranked Data Aggregation (TopKLists)

By pooling findings across different datasets, meta-analysis helps in identifying patterns, biomarkers, and pathways that remain consistent across various experiments. This is particularly valuable in cancer omics research, where high costs of experiments often necessitate combining data from multiple, diverse datasets to achieve reliable sample sizes. In this context, rank-based approaches play a particularly significant role. One reason is that omics data often come from different platforms, each with unique technical characteristics and distributions. Absolute measurements from one platform may not be directly comparable to another. Ranked approaches solve this problem by focusing on the relative ordering of features instead of their absolute values, reducing biases caused by platform-specific differences.

This was a driver of our research where we developed methods for the statistical inference and aggregation of ranked omics datasets, which led to the development of the TopKLists R package [2]. This package was specifically designed to tackle the challenges of integrating data from different high-throughput platforms, where datasets often vary in list lengths, measurement techniques, and even the items being ranked. By focusing on ranked lists, TopKLists provides a way to consolidate platform-independent results, making it particularly relevant for omics research.

The package includes three main modules. TopKInference estimates the optimal length of top-k lists for integration, even in noisy or incomplete rankings. It uses a moderate deviation-based method to handle cases where the reliability of rankings decreases after the first k items due to technical or biological variability. TopKSpace then aggregates these top-k lists using algorithms like Borda's method, Markov chain techniques, and a more precise cross-entropy Monte Carlo (CEMC) method. These approaches consider weighted distances, such as Kendall's τ or Spearman's footrule, to create a consensus ranking. Finally, TopKGraphics provides graphical tools to explore and visualize ranked data, helping users interpret results and select parameters, such as with the Δ-plot for visualizing inter-platform variability. The package has been optimized for use on standard desktop computers, with computationally heavy sections implemented in C to speed up processing. Most tasks, even for rankings of thousands of items, are completed in seconds, although the stochastic aggregation methods can take slightly longer. A graphical user interface (GUI) has been developed for TopKLists, using the gWidgets2 package, which makes it more accessible for users without advanced programming skills. TopKLists is freely available under the LGPL-3 license and can be downloaded from CRAN, with additional resources, documentation, and the latest development version available on its R-Forge page [13]. The package has already been applied to integrate microRNA data from non-small cell lung cancer studies conducted on multiple platforms, demonstrating its practical utility in handling ranked data from diverse sources.

Topological Pathway Analysis (ToPASeq)

Pathway analysis is a crucial step in interpreting results from molecular analyses, providing a biological context for the observed changes in gene or protein expression. By mapping these changes onto known biological pathways, researchers can uncover mechanisms underlying differences between conditions, disease progression, or other clinical outcomes. This type of analysis is often the logical next step in exploratory studies, transforming lists of differentially expressed genes into meaningful insights about cellular processes.

There are two main approaches to pathway analysis: overrepresentation analysis (ORA) and topology-based methods. ORA identifies pathways that are significantly enriched with genes of interest, without considering their interactions or positions within the pathway. While straightforward, ORA assumes that all genes in a pathway are equally important, potentially missing key insights. In contrast, topology-based approaches account for the structure of the pathway, incorporating information about gene positions, interactions, and roles within the network. This additional layer of context allows researchers to prioritize biologically meaningful changes and identify key regulatory nodes, which are often critical for understanding disease mechanisms. By leveraging topology, these methods provide a more accurate and nuanced understanding of the biological processes at play, making them particularly valuable for studies aiming to uncover the mechanisms driving colorectal cancer heterogeneity or progression.

The landscape of topology-based pathway analysis methods is highly diverse, with each method employing distinct frameworks and assumptions to interpret molecular data. These differences often result in significant variability in the pathways identified as relevant, making it challenging for researchers to determine the most appropriate tool for their specific datasets and research questions. Recognizing this, we conducted a comprehensive comparison of seven representative topology-based pathway analysis methods [***4***]. Our aim was to evaluate their strengths and limitations across multiple criteria, guiding researchers in selecting the optimal method for their studies.

To support this work and facilitate the practical application of topology-based pathway analysis, we developed a new R/Bioconductor package, ToPASeq [***3***]. This package offers a uniform interface to the seven analyzed methods, three of which we implemented *de novo* and four adapted from existing implementations. ToPASeq also includes tailored visualization tools, as well as functions for importing and manipulating pathways and their topologies, enabling its application across various species. The package is designed to analyze differential expressions of pathways between two conditions and is compatible with both gene expression microarray and RNA-Seq data. Written in R and distributed under an AGPL-3 license, ToPASeq is freely available from Bioconductor 3.12 [14,15], providing the research community with a powerful and accessible toolkit for pathway analysis.

The comparison was based on an extensive set of criteria, addressing both dataset characteristics and methodological aspects. Data set-centric parameters included sample size, pathway size, the number of differentially expressed genes (DEGs) in the dataset, and thresholds used to identify DEGs. These factors were tested to describe the performance of each method under various conditions and provide recommendations for selecting the best tool for specific dataset configurations. In addition, the ability of the methods to control type I error was evaluated, ensuring reliability when no true signal exists. We also examined how the methods handled specific biological and technical challenges. For example, we tested the influence of overexpression of individual genes, the discarding of topological information, and the preprocessing of pathway topologies. These experiments were critical to assess whether the methods genuinely leveraged pathway topology in their analysis. If no effects were observed under these conditions, the method could not be considered a true topology-based approach. Furthermore, we evaluated the increased sensitivity and specificity expected from incorporating topological information by assessing the identification of biologically relevant pathways, which is crucial for advancing our understanding of molecular mechanisms.

Additional computational tools for integrating molecular data with image analysis were developed, and these will be discussed in detail in chapter 3.4.

3.2. MOLECULAR SUBTYPING AND TUMOR HETEROGENEITY

The last fifteen to twenty years have seen an intensive search for molecular markers of cancer progression and a deeper understanding of the biology underlying (non)-response to therapy. Colorectal cancer (CRC) is no exception, with many significant discoveries advancing our knowledge of the heterogeneity of this disease. The common approach to diagnosing CRC relies on a combination of clinical evaluation, endoscopic examination, and histopathological analysis of biopsy samples. Standard treatment strategies include surgical resection, often followed by adjuvant chemotherapy, particularly for stage III and high-risk stage II cases. Targeted therapies, such as those inhibiting EGFR or VEGF pathways, are used in advanced disease based on molecular profiling.

In diagnostics, clinical and histopathological markers play a key role. Staging based on the TNM system (Tumor, Node, Metastasis) remains the cornerstone for assessing disease severity and guiding treatment decisions. Histopathological features such as tumor grade, lymphovascular invasion, and presence of perineural invasion provide additional prognostic information. Another critical factor in CRC is tumor sidedness, which reflects distinct biological and clinical differences between tumors originating in the proximal (right-sided) and distal (left-sided) colon. These differences are partly attributed to the embryonic development of the gut, where the right side arises from the midgut and the left side from the hindgut. Additionally, as I will discuss in chapter 3.5, this variation is likely influenced by the site-specific microbial composition of the gut.

Aside from histopathological features and staging, molecular testing has become increasingly important in CRC diagnostics, particularly for identifying mutations with therapeutic implications. The most commonly tested mutations include those in the *KRAS* and *NRAS* genes, as their presence predicts resistance to anti-EGFR therapies. *BRAF* mutations, particularly the V600E variant, are associated with a poor prognosis and also influence treatment strategies. Additionally, testing for mismatch repair (MMR) deficiency or microsatellite instability (MSI) is now standard, as these biomarkers can identify patients eligible for immune checkpoint inhibitors.

This section explores the identification and characterization of molecular subtypes and heterogeneity in colorectal cancer (CRC), emphasizing their clinical implications. Understanding CRC heterogeneity at the molecular, genetic, and structural levels is vital for refining therapeutic strategies and improving patient outcomes. My journey in this field began with the unique opportunity to contribute to the analysis of molecular data from the PETACC-3 clinical trial [16]. The PETACC-3 clinical trial provided a unique opportunity to analyze colorectal cancer (CRC) at multiple molecular levels, combining transcriptomics, comparative genomic hybridization (CGH), histopathological images, and clinical molecular markers with comprehensive clinical data, including long-term follow-up for prognosis modeling. This experience led to my long-term interest in the subject and has shaped my scientific research career.

Supervised approach to molecular profiling in CRC: Insights from Tumor Location, Mutational Status, Transcriptomics and Copy Number Aberrations

Leveraging the unique PETACC-3 dataset, we performed comprehensive analyses to explore how tumor location (proximal vs. distal) and mutational status and morphology influence molecular profiles, clinical parameters, patient prognosis and response to therapy.

In [***5***] we conducted a comprehensive analysis of somatic copy number aberrations (CNAs) in 302 stage II/III CRC samples from PETACC-3. The aim was to provide a detailed molecular overview of CNAs, elucidate their underlying biology, and explore associations with clinical outcomes. We identified regions of recurrent CNAs, comprising both well-established oncogenes (e.g. *MYC*, *EGFR*, and *CCND1*), as well as novel loci with potential biological significance. Notably, amplification of 12p13.33 revealed *WNK1* as a candidate oncogene implicated in MAPK signaling and various cancer hallmarks, while multiple loci on 20q (including 20q11.21, 20q13.12, and 20q13.31) pointed to oncogenic drivers such as *HNF4A, WISP2,* and *BMP7*, which are involved in epithelial-mesenchymal transition, metastasis, and tumor aggressiveness. Additionally, our findings on 10p15.3-p14 and 19p13.12 deletions linked these loci to poor survival outcomes, while 20q gains were unexpectedly associated with better overall survival in stage III tumors, contrasting prior reports. This was later shown to be consistent with definition of CRC molecular subtypes [***9***], see below. Importantly, our study highlighted a novel aspect of CNA interactions: significant non-random correlations between unlinked DNA loci. This observation brought hypothesis of emergence of highly ordered structural changes during tumor progression, potentially driven by selective pressures acting on tumorigenic pathways.

Right-sided CRCs often exhibit features such as microsatellite instability, higher mutation burden, and immune infiltration, and are associated with a worse prognosis and reduced response to anti-EGFR therapies compared to left-sided tumors. In contrast, left-sided CRCs are typically chromosomally unstable and show better responses to targeted therapies, highlighting the need to consider tumor location in treatment planning. Our work in [*6*] was among the studies that contributed to this understanding, offering key insights into the molecular and clinical differences between right- and left-sided CRCs by leveraging data from the PETACC-3 trial. We confirmed the well-established observation that proximal tumors are more frequently microsatellite unstable (MSI) and hypermutated, largely due to deficiencies in DNA mismatch repair (MMR). Even among microsatellite stable (MSS) proximal tumors, we found an enrichment of potentially deleterious mutations, including alterations in *KRAS*, *BRAF*, and *PIK3CA*. Consistent with prior studies, we observed that proximal tumors are often mucinous, densely infiltrated by tumor-infiltrating lymphocytes, and exhibit activated MAPK signaling. They also frequently express a serrated pathway signature and a high BRAF score, indicating pathway activation even in the absence of *BRAF* mutations. Potential contributors to these features are side-specific environmental factors (e.g., bacterial toxins, mutagenic metabolites) and tolerance to DNA repair defects and oncogenic stress. For distal CRCs, our work corroborated the frequent presence of large-scale chromosomal alterations, including 18q loss and 20q gain (leveraging data from [*5*]), hallmark features of chromosomal instability. We also observed the activation of EGFR signaling, with *HER1* and *HER2* amplifications present in a subset of distal tumors, particularly those wild-type for *KRAS* and *BRAF*. These findings suggested the importance of EGFR pathway activation in distal colon carcinogenesis and its potential as a therapeutic target. Beyond confirming these previously known patterns, our analysis provided new insights into the relationship between tumor location and clinical outcomes. We showed that tumor location acts as an independent prognostic factor for survival after resection (SAR) and relapse-free survival (RFS). Proximal tumors, even when MSS, were associated with higher mutation rates and cellular plasticity, which may exacerbate the deleterious effects of chemotherapy. These features likely contribute to poorer outcomes under current treatment regimens, suggesting that proximal tumors may require entirely different therapeutic approaches. Our observations also reinforced the benefit of anti-EGFR therapies in distal CRCs. We found that EGFR pathway activation in distal tumors makes them more responsive to anti-EGFR agents than proximal tumors, later evidenced by results from a clinical study [17].

Another influential result in this category was the identification of a subgroup of colorectal tumors with a *BRAF* wild-type (*BRAFm*-like) phenotype but molecular profiles resembling *BRAF*-mutated (*BRAFm*) tumors [7]. This subgroup was identified through a high-sensitivity gene expression signature derived from *BRAFm* tumors, which was robust enough to support a patent filing for the *BRAFm*-like signature [18]. The methodology for developing this classifier was implemented using the Rgtsp tool, as previously described in chapter 3.1. The *BRAFm*-like subgroup was found to also share clinicopathologic features with *BRAFm* tumors, such as enrichment for MSI-H, mucinous histology, and right-sided location. Frequencies of high-grade tumors were 30% in *BRAFm*, 20% in *BRAFm*-like, and only 5% in predicted *BRAF* wild-type (pred-*BRAFwt*) tumors, while MSI-H rates were 30%, 30%, and 3%, respectively. Interestingly, this group also showed poor prognosis, even in microsatellite stable (MSS) cases (Figure 1). Importantly, this finding challenged the conventional understanding that *KRAS*-mutated tumors form a homogeneous group, as the *BRAFm*-like subgroup included part of tumors with *KRAS* mutations as well as double wild-type (WT2) samples. Additionally, *BRAFm*-like tumors demonstrated a distinct adenoma-carcinoma progression sequence linked to the serrated pathway, suggesting a shared underlying biology with *BRAFm* tumors. From a biological perspective, the *BRAFm*-like subgroup highlights tissue-specific biology in CRC compared to melanoma, where *BRAFm* inhibitors have been successful. This tissue-specific biology may explain why inhibitors like PLX4032, effective in *BRAFm* melanoma, have shown limited efficacy in *BRAFm* CRC. The study's results underscored the need for a revised definition of CRC subgroups, particularly within *KRAS*-mutated tumors, and provided a framework for developing tailored therapeutic strategies.

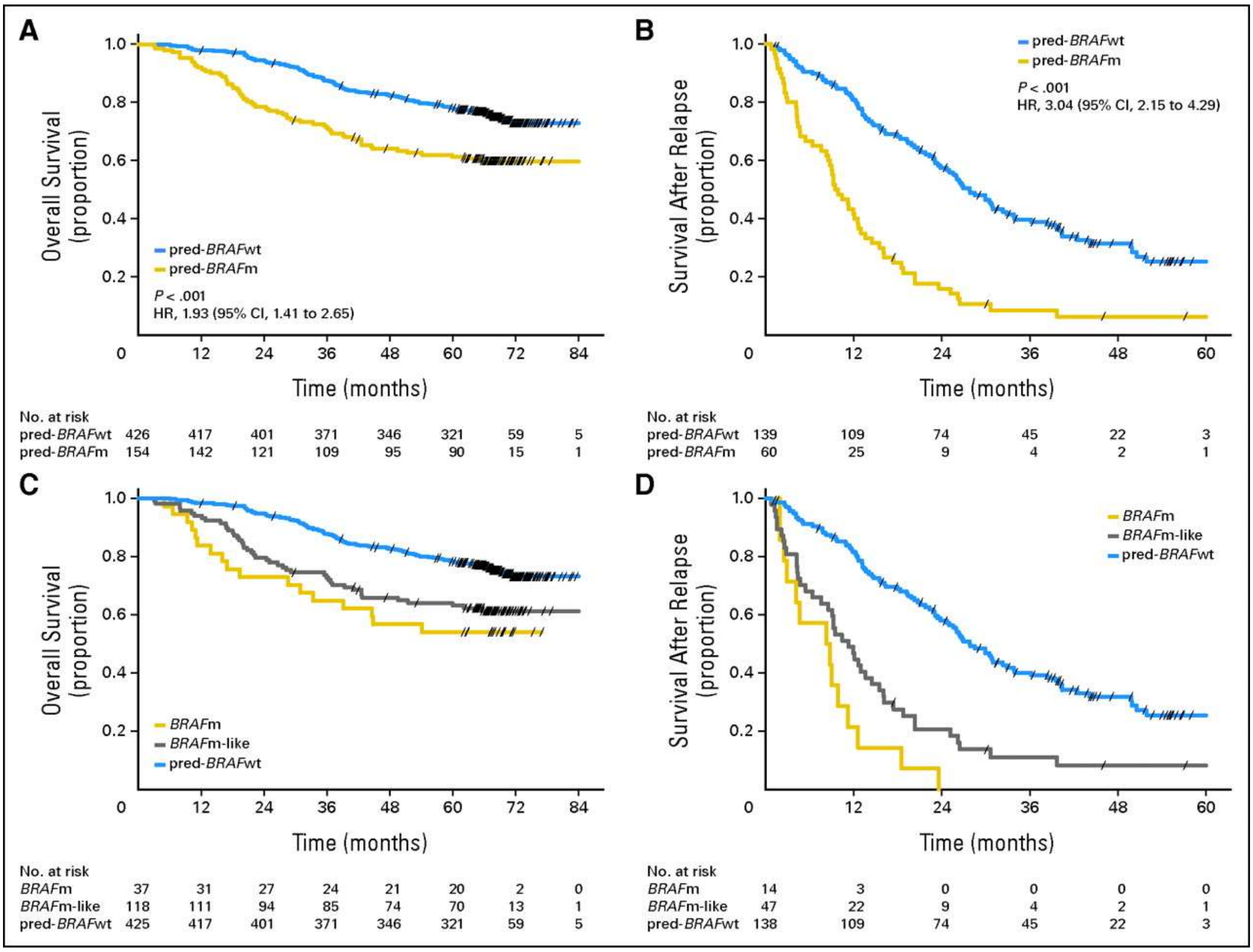


Figure 1. Kaplan-Meier curves for different stratifications of the stage III subpopulation and different end points. Columns correspond to overall survival and survival after relapse end points, respectively. Panels A-D correspond to stratifications into samples predicted to be BRAF mutant (pred-BRAFm)/predicted to be BRAF wild type (pred-BRAFwt; A, B) and BRAF mutant (BRAFm)/BRAF mutant like (BRAFm-like)/pred-BRAFwt (C, D) in the whole stage III subpopulation (from [7])

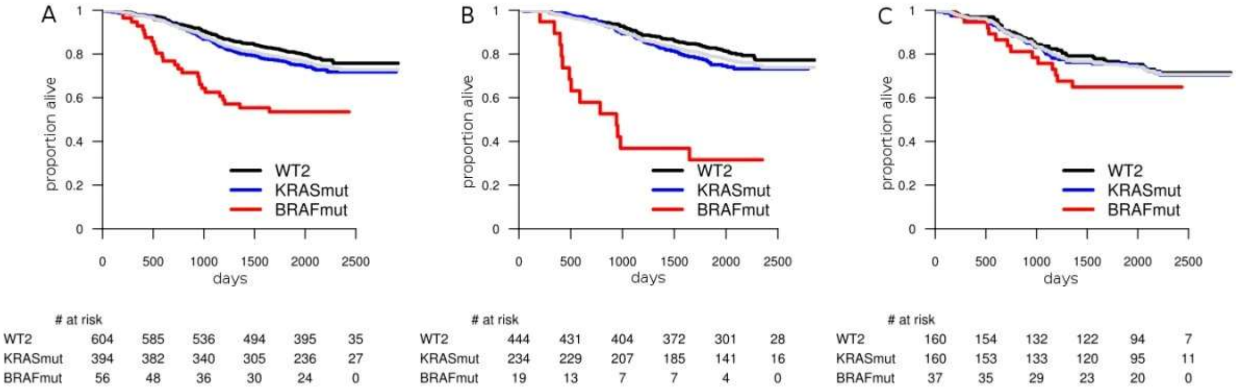
In this work we established a novel subgroup with clear prognostic and histological significance and demonstrated the value of gene expression profiling in refining CRC classification. It also supported the need for further functional investigations and clinical trials aimed at identifying actionable targets within the *BRAFm*-like population and subsequent functional investigations, including search for actionable targets [19] and clinical trials [20].

Context-Dependent Prognostic Value of BRAF and KRAS Mutations

While *BRAF* and *KRAS* mutations have been widely studied as prognostic markers in CRC, their predictive value has remained controversial, particularly for *KRAS*. The prevailing assumption in earlier studies was that these mutations have a uniform prognostic effect across all patients, independent of clinical context. However, our work in [8] demonstrated that the prognostic impact of *BRAF* and *KRAS* mutations is highly context-dependent, varying significantly based on tumor location and microsatellite instability (MSI) status.

By leveraging the PETACC-3 dataset, which included mutation data from over 1,400 stage II–III CRC patients, we systematically assessed the prognostic value of *BRAF* and *KRAS* mutations across multiple clinically relevant subgroups. To ensure statistical robustness, only subgroups with at least 20 patients were considered for prognostic assessment. We employed univariate survival analyses using log-rank tests and estimated hazard ratios (HR) for overall survival (OS), relapse-free survival (RFS), and survival after relapse (SAR). Multiple testing correction was applied using the Bonferroni method, setting a stringent significance threshold (adjusted $p \leq 0.01$). To assess potential interactions, we further performed multivariate Cox regression models incorporating second-degree interaction terms between MSI status, *BRAF* mutation, and tumor location, adjusting for grade, T stage, and N stage.

Our analysis confirmed that *BRAF* mutations were strongly prognostic for overall survival (OS) and survival after relapse (SAR), but notably, this effect was almost entirely driven by microsatellite stable (MSS) tumors located in the left colon. Within this subgroup, *BRAF* mutations conferred a six-fold increase in mortality risk compared to *BRAF*-wild-type MSS left-sided tumors, whereas in MSI-high (MSI-H) or right-sided tumors, *BRAF* mutations had no significant prognostic value (Figure 2). This observation challenged the widespread practice of reporting hazard ratios for BRAF mutation without considering tumor location and MSI status, highlighting the need for more nuanced interpretation of prognostic biomarkers in CRC.



# at risk						
WT2	604	585	536	494	395	35
KRASmut	394	382	340	305	236	27
BRAFmut	56	48	36	30	24	0

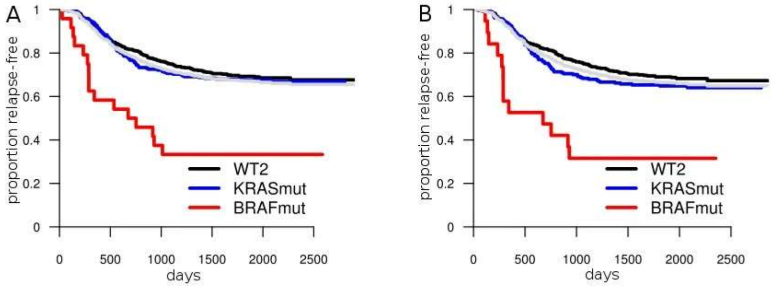
# at risk						
WT2	444	431	404	372	301	28
KRASmut	234	229	207	185	141	16
BRAFmut	19	13	7	7	4	0

# at risk						
WT2	160	154	132	122	94	7
KRASmut	160	153	133	120	95	11
BRAFmut	37	35	29	23	20	0

Figure 2. Overall survival: prognostic value of BRAF and KRAS mutations within MSS and by tumor site. A: all MSS tumors; B: MSS left-sided tumors; C: MSS right-sided tumors. The light gray survival curve represents the whole subpopulation survival (A: all MSS, B: MSS left-sided, C: MSS right-sided tumors) (from [8])

For relapse-free survival (RFS), we made the novel observation that *BRAF* mutations were also prognostic in MSS left-sided tumors, contradicting prior studies that did not find an association between *BRAF* and relapse (Figure 3). Importantly, these results were validated in multivariate models that accounted for tumor grade, T stage, and N stage, reinforcing the robustness of the findings.

In contrast, *KRAS* mutations did not reach statistical significance as a prognostic marker for OS, SAR, or RFS in the overall cohort. However, our stratified analyses revealed that *KRAS* mutations showed trends towards significance in certain subpopulations, particularly for RFS in right-sided tumors. Intriguingly, in MSI-H right-sided tumors, *KRAS* mutations appeared to have a protective effect, identifying a subset of patients with better prognosis. While these results did not reach the stringent significance threshold after multiple testing correction, they suggest that the prognostic role of KRAS may be more complex than previously assumed. Our findings support the hypothesis that the KRAS-mutant population is molecularly heterogeneous, which may explain the inconsistent prognostic associations reported in the literature.



# at risk						
WT2	493	421	369	339	282	28
KRASmut	266	225	191	179	150	19
BRAFmut	24	14	9	8	6	1

# at risk						
WT2	444	378	331	303	255	26
KRASmut	234	196	163	151	127	15
BRAFmut	19	10	6	6	4	0

Figure 3. Relapse-free survival: prognostic value of BRAF and KRAS in left-sided tumors. A: all left-sided tumors; B: MSS left-sided tumors. The light gray survival curve represents the whole subpopulation survival (A: all left tumors; B: MSS left).

These results provided a key conceptual advance: the prognostic value of oncogenic mutations in CRC cannot be interpreted in isolation but must be considered within the broader tumor context. This insight has direct implications for clinical trial design, biomarker interpretation, and the development of prognostic gene signatures, reinforcing the need for stratification by MSI status and tumor location in future studies.

Unsupervised approach to CRC molecular heterogeneity – the CRC molecular subtypes

While supervised approaches to exploring tumor heterogeneity are informative, they are inherently limited in their ability to uncover the unknown. Unsupervised methods, such as clustering, address critical questions like: 'Do we have enough information? Are we overlooking key insights? What if our existing classifications are incorrect?'. In 2015, an important study introduced the Consensus Molecular Subtypes (CMS), a framework that stratifies CRC into biologically distinct groups based solely on gene expression profiles of tumors, revealing subtype-specific differences in prognosis and therapy response. Our work in [9] played a pivotal role in the development of the CMS framework. Notably, our group was among the first to initiate efforts to define molecular subtypes of CRC. Our subtyping system was one of the six approaches included in the CMS study, reflecting its significance in shaping the consensus. Furthermore, the robustness of our subtypes was evident in their strong alignment with the CMS subtypes, highlighting the reproducibility as well as biological validity of our methodology (Figure 4).

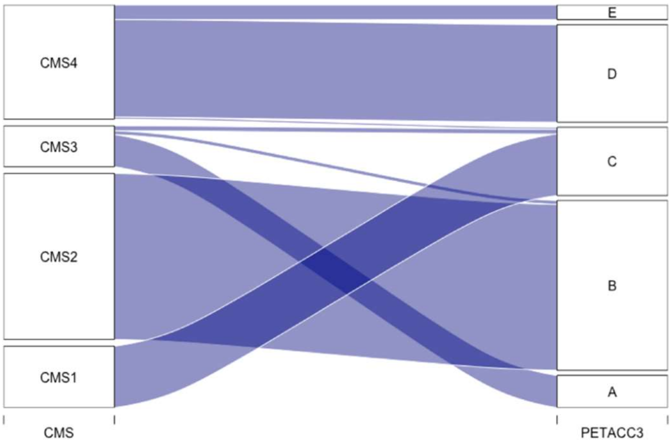


Figure 4. Sankey diagram of concordance between five Budinska gene expression subtypes (right) [9] and 4 CMS subtypes (left).

The methodological approach we employed for gene expression-based subtyping was designed to ensure robustness and biological relevance. It consisted of the following key steps:

i) Dimensionality Reduction:

To reduce noise and focus on informative features, we implemented a multi-step dimensionality reduction process. First, we filtered for genes with high coefficients of variation (CV), eliminating low-expressed and non-informative genes. Next, we grouped the remaining genes into clusters based on the correlation of their expression patterns, summarizing each cluster as a meta-gene represented by the median expression value of its member genes. These meta-genes were further clustered into higher-order structures, which were then subjected to gene set enrichment analysis. This final step enabled the identification of key biological processes driving CRC heterogeneity.

ii) Consensus Clustering and Dynamic Hybrid Tree Cut for Sample Stratification:

We applied robust consensus clustering [21] to group samples based on their meta-gene expression profiles into five distinct clusters. The principle of consensus clustering lies in aggregating results from multiple clustering iterations (in our case, hierarchical clustering), typically using subsampled data, to identify stable and reproducible groupings. This approach mitigates the sensitivity to initial conditions inherent in many clustering methods, leading to more reliable and reproducible results. This robustness is particularly relevant for high-throughput molecular datasets, which are often prone to technical variability and batch effects.

In each iteration, clusters were determined using a dynamic tree cut procedure [22], which provided a more robust and adaptive method for defining cluster boundaries. The dynamic hybrid method offers significant advantages over traditional fixed-height cutoffs for hierarchical clustering, particularly in the context of CRC molecular subtyping. Unlike fixed methods, it adapts to the shape and structure of dendrogram branches, enabling the detection of clusters with varying sizes and densities, which is crucial for capturing CRC heterogeneity. Additionally, it identifies nested clusters and effectively handles outliers, ensuring that subtle subtypes and unique molecular signatures are not misclassified or lost. Its flexibility allows for parameter tuning and automation, making it well-suited for large-scale genomic datasets. These features make the dynamic hybrid method a robust and precise tool for defining clinically relevant CRC subtypes. The final optimal number of clusters was determined using the consensus index, ensuring statistical robustness. To maintain homogeneity within the identified groups, we excluded samples that did not cluster with high probability. These outliers may represent rare subtypes or technical artifacts and excluding them helped to prevent the distortion of group-specific characteristics.

iii) Validation Across Independent Datasets:

The robustness of our findings was further validated by performing clustering across five independent external datasets. Cross-cluster training of classifiers was employed to assign samples to their respective groups in these external cohorts, ensuring the reproducibility of our methodology. We compared the molecular, prognostic, mutational, and clinical/histopathological features of the groups across datasets, confirming the consistency and biological relevance of the identified subtypes.

This approach resulted in the generation of five CRC molecular subtypes (A-E), characterized by distinct molecular processes, mutation profiles, and differences in overall survival (OS), relapse-free survival (RFS), and survival after relapse (SAR) (Figure 5). Importantly, and in contrast to other groups [23–26], we also characterized these subtypes from a histopathological perspective. Initially, our expert pathologist performed an unsupervised assessment of common histopathological features, such as tumor budding, hypoxia, peritumoral lymphocytic infiltration, necrosis etc. Statistical analysis, however, revealed no significant differences in the distribution of these characteristics among the subtypes. In a subsequent supervised analysis, we identified that the proportion of distinct morphologies - namely complex tubular, mucinous, solid/trabecular, serrated, papillary, and desmoplastic - varied significantly between subtypes. This finding aligned with molecular data.

Subtype A, which exhibited a gene expression signature associated with differentiated colon, referred to as *surface-crypt-like*, was notably enriched in papillary and serrated morphologies. Papillary adenocarcinomas are characterized by finger-like epithelial projections, reflecting differentiation patterns akin to normal colonic crypts. Serrated adenocarcinomas, with their hallmark saw-toothed glandular pattern, arise from serrated polyps following the serrated neoplasia pathway, which involves BRAF mutations and Wnt signaling pathway alterations. These morphologies align with the differentiation process of normal colonic crypts, where stem cells at the crypt base give rise to epithelial cell types that migrate and differentiate along the crypt axis. Subtype B - *Lower-crypt lik*e, showed enrichment in the complex tubular pattern, aligning with the typical morphology of colorectal adenocarcinoma. This subtype's molecular profile exhibited active proliferation, amplification of chromosomes 20q and 20p, and deregulation of genes commonly associated with intestinal differentiation, including *CDX2, IHH, VAV3, ASCL2,* and *PLAGL2.* Histologically, this subtype demonstrated low immune cell infiltration and a minimal presence of epithelial-mesenchymal transition (EMT) or stromal components, consistent with a more proliferative and less invasive phenotype. Immunohistochemical staining revealed active β-catenin signaling, a hallmark of the Wnt pathway, which is frequently implicated in colorectal carcinogenesis and reflects the molecular processes driving tumor progression in this subtype. *Subtype C (CIMP-H like)* displayed molecular and clinicopathological characteristics that closely align with the well-established CIMP-H phenotype of colorectal cancer. This subtype expressed the BRAF-mutant signature identified in our earlier work [7] (87.0% of cases) and a robust CIMP-H signature, characterized by widespread promoter hypermethylation. In addition, subtype C was enriched for MSI, right-sided location, and mucinous histology, hallmarks of the CIMP-H phenotype. Similar to previously reported hypermutated tumors, this subtype exhibited a low frequency of copy number variations (CNVs), suggesting that its tumorigenesis is driven more by epigenetic and mutational events than by chromosomal instability.

Most interesting, however, was Subtype D, which exhibited a molecular signature strongly indicative of epithelial-to-mesenchymal transition (EMT), characterized by high immune cell infiltration and low proliferation. This suggested that the tumor might comprise cancer cells actively undergoing the process of EMT. Surprisingly, histopathological examination revealed that this molecular profile was not primarily due to high proportion of mesenchymal tumour cells but rather result of a high desmoplastic reaction. A desmoplastic reaction refers to the excessive growth of fibrous or connective tissue, often triggered by interactions between tumor cells and the surrounding stromal microenvironment. This process creates a dense, fibrotic stroma composed of activated fibroblasts, immune cells, and extracellular matrix proteins, which can mimic EMT at the molecular level by inducing similar gene expression patterns.

Interestingly, Subtype D exhibited significantly lower overall survival (OS) and relapse-free survival (RFS), even after accounting for other clinically relevant variables, such as tumor stage or MSI status. The association between tumour stromal content and prognosis was not entirely novel [27], but our identification of stromal enrichment within unsupervised molecular subtypes prompted further studies to explore this relationship in greater depth [28]. Summary of subtype characteristics is shown in Table 1.

This work consequently steered my research path in a very specific direction, focusing on the integration of histopathological and molecular data to better understand tumor heterogeneity in CRC.

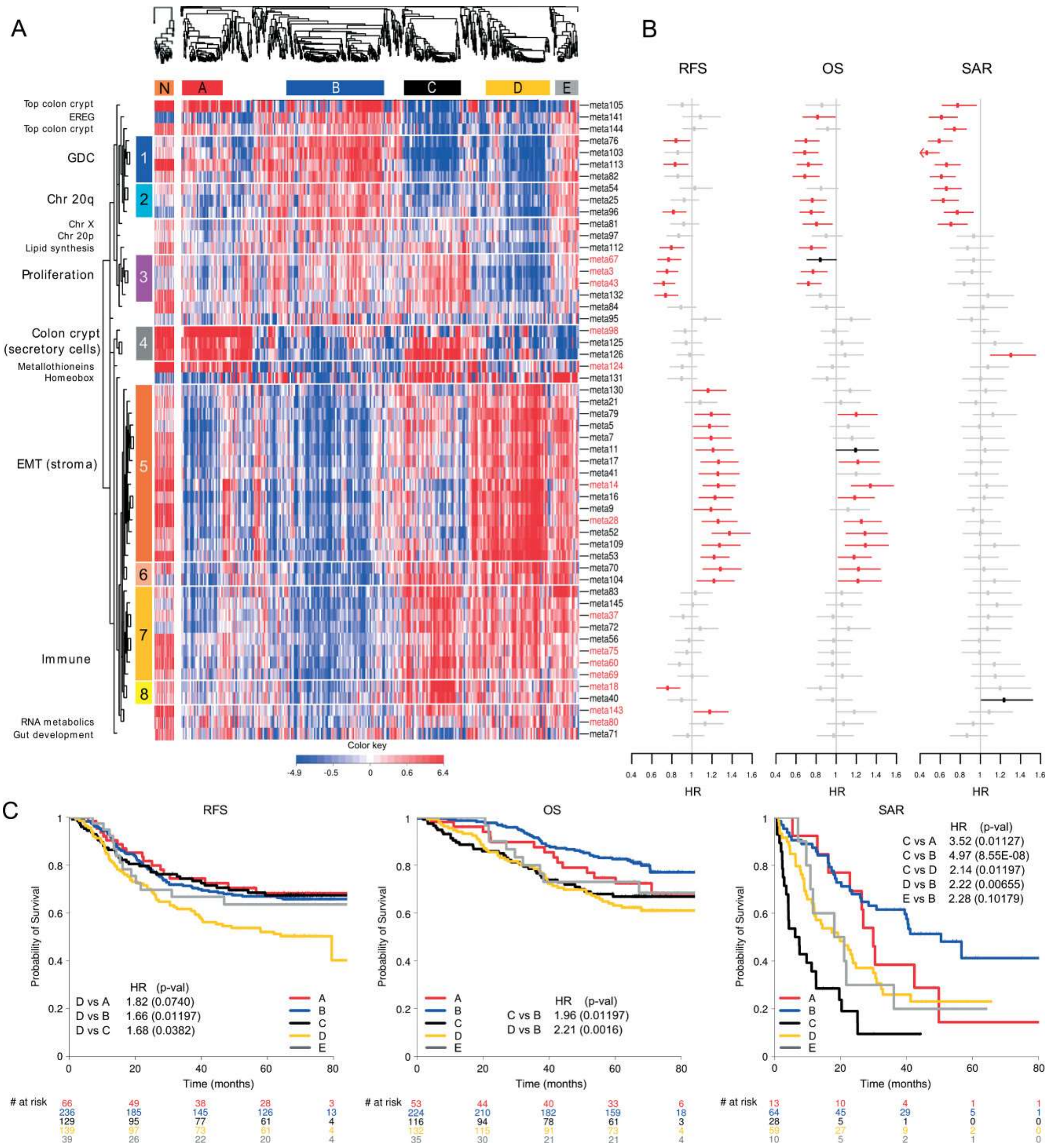


Figure 5. Meta-gene expression pattern in subtypes, connected with prognostic effect of subtypes and meta-genes, in the discovery set. (A) Two heat maps clustering normal (left) and CRC (right) samples (columns) and meta-genes (rows). Colours represent decreased (blue) or increased (red) meta-gene expression relative to their medians. Normal samples were clustered independently on meta-genes centred to CRC meta-gene medians. For comparative purposes, ordering of meta-genes in normal samples is imposed to correspond to that of CRC samples. White horizontal lines denote eight unsupervised clusters of meta-genes, each assigned a colour bar on the left; meta-genes not belonging to a cluster have no colour bar. Names of the meta-genes corresponding to gene modules with gene–gene correlations in normal samples comparable to those in cancer samples are marked red (see Supplementary material, Figure S1D). (B) Effect of inter-quartile range (IQR) standardized expression of meta-genes on RFS, OS and SAR. Points represent estimated hazard ratio (HR), bars represent 95% CI. Bold lines represent effects significant at 5% without adjustment for multiple hypothesis testing; red lines represent effects significant at FDR < 10%; details are provided in Table S6 (see Supplementary material). (C) Kaplan–Meier plots for RFS, OS and SAR, with HR for significant pairwise comparisons (*p* values adjusted for FDR). Numbers below *x* axes represent number of patients at risk at selected time points. (from [*9*])

Table 1. Summary of subtype characteristics (adjusted from [*9*])

	Molecular subtypes				
Feature	**A: Surface crypt-like**	**B: Lower crypt-like**	**C: CIMP-H-like**	**D: Mesenchymal**	**E: Mixed**
MSI		–	+		–
BRAF	+	–	+		–
KRAS	–				
P53			–		+
Histopathology	Papillary or serrated	Complex tubular	Solid/trabecular or mucinous	Desmoplastic	Complex tubular
IHC (Nuclear β-catenin at IF)	–	+	–	–	+
Median Survival (months)	NA (OS), NA (RFS), 28.9 (SAR)	NA (OS), NA (RFS), 50.4 (SAR)	NA (OS), NA (RFS), 6.9	NA (OS), 79.5 (RFS), 19.8 (SAR)	NA (OS), NA (RFS), 19.6 (SAR)
Clinical Site		Left	Right		Left
Grade		2	3		
Up-regulated Genes	Top colon crypt, secretory cell, metallothioneins	Top colon crypt, proliferation, Wnt	Proliferation, immune, metallothioneins	EMT/stroma, CSC, immune	EMT/stroma, immune, top colon crypt, Chr20q, GDC, CSC
Down-regulated Genes	EMT/stroma, Wnt, CSC, Chr20q, proliferation	EMT/stroma, immune, secretory cell	GDC, top colon crypt, Chr20q	Proliferation, secretory cell, top colon crypt, GDC, Wnt, Chr20q	Secretory cell

3.3. PRECLINICAL MODELS

The study of colorectal cancer (CRC) has been transformed by the integration of computational modeling with experimental validation. While bioinformatics-driven analyses reveal critical aspects of tumor heterogeneity, drug resistance, and molecular subtypes, their true impact lies in biological validation. Preclinical models serve as a crucial link, allowing us to test hypotheses, confirm computationally derived biomarkers, and dissect the mechanisms underlying tumor progression and therapy response.

Different experimental models offer complementary advantages, each addressing distinct aspects of CRC biology. Genetically engineered mouse models (GEMMs) provide a controlled system to study oncogenic pathways in vivo, capturing key molecular hallmarks of CRC. Patient-derived xenografts (PDXs) have proven invaluable for precision oncology, faithfully preserving patient-specific tumor characteristics and drug response profiles. More recently, patient-derived organoids (PDOs) have emerged as a versatile ex vivo system, bridging the gap between in vitro experimentation and in vivo validation, enabling high-throughput functional studies on molecular subtypes and therapy resistance. Beyond their individual strengths, preclinical models are most powerful when combined with computational analyses. Cross-species transcriptomic comparisons refine computational predictions by identifying conserved molecular programs, while functional studies in PDX and organoid models validate key molecular drivers and therapeutic targets. These models have also provided new perspectives on tumor-microenvironment interactions, immune infiltration, and the role of the microbiome in CRC progression.

Cross-Species Transcriptomic Analysis of CRC: Insights from Genetically Engineered Mouse Models

For decades, GEMMs have been instrumental in modeling human CRC by introducing mutations in key driver genes, such as *APC, TP53, KRAS,* and *BRAF*, which are frequently altered in human tumors [29]. Unlike traditional cancer models, GEMMs allow for the stochastic and tissue-specific activation of these mutations, mimicking the sporadic nature of human CRC development. By combining GEMMs with high-throughput transcriptomic profiling, it is possible to assess how specific genetic alterations shape the tumor microenvironment and contribute to disease progression. I had the unique opportunity to be involved in such efforts, where our expertise in mining large-scale CRC transcriptomic datasets was applied to evaluate the molecular fidelity of GEMMs and assess their relevance to human disease [***10***]. This study focused on establishing genotype-specific gene expression signatures in GEMMs and determining their molecular resemblance to human CRC and their utility in preclinical research. Gene expression profiling of GEMM-derived tumors was performed, and mutation-specific transcriptional signatures were identified through multivariable statistical modeling. These signatures were then validated in clinically annotated human CRC datasets (PETACC-3, GSE14333), revealing a strong overlap between the GEMM *KRAS* signature and human *KRAS*-mutant and *BRAF*-like tumors, both of which are associated with poor prognosis and MAPK pathway activation. In contrast, the *BRAF* signature did not align well with human *BRAF*-mutant CRC, likely reflecting biological differences in APC co-mutation frequencies. Further, the GEMM KRAS signature predicted increased sensitivity to MEK inhibitors (PD-0325901, AZD6244) in CRC cell lines, providing a potential tool for therapeutic stratification. This confirmed the relevance of GEMMs in modeling CRC heterogeneity and emphasized the need for refined models to better capture BRAF-driven disease biology.

Data Integration Challenges in PDX Models: Bridging Preclinical and Clinical Insights

Patient-derived xenografts (PDXs) have emerged as a powerful tool in translational oncology, enabling high-throughput studies that link genetic and functional characteristics to therapeutic responses. However, the large-scale use of PDX models presents significant challenges, particularly in data management, integration, and analysis. The central focus of my work was developing strategies to harmonize preclinical PDX data with molecular classifications derived from patient tumors. This involved addressing the biological variability introduced during tumor engraftment and propagation, as well as ensuring robust data normalization, standardization of sample metadata, and applying analytical corrections to account for systematic biases, such as the loss of human immune and stromal components. This expertise was integrated into the review by Byrne et al. (2017) [***11***], where we critically assessed the role of PDXs in cancer precision medicine, highlighting both their advantages and limitations in preclinical research. The review examined how PDX models can bridge the gap between laboratory findings and clinical applications, especially in drug development and biomarker discovery. We emphasized the importance of standardized protocols, rigorous data integration, and careful result interpretation to maximize the translational value of PDX models in oncology. My contribution to the review focused specifically on the complexities of data stratification in

PDX studies, where we proposed computational solutions to mitigate population selection biases and improve integrative analyses across various experimental platforms. In particular, I helped design analytical workflows to standardize these processes, ensuring that PDX models could be meaningfully aligned with clinically relevant subgroups. These efforts are essential for enhancing the translational potential of PDX-based approaches, particularly in preclinical drug testing and biomarker discovery.

3.4. INTEGRATING DIGITAL PATHOLOGY AND OMICS DATA

Histopathological evaluation has long been a cornerstone of cancer diagnostics, providing essential insights into the structural and cellular organization of tumors. By examining stained tissue slides under the microscope, pathologists assess key features such as tumor grade, cellular morphology, tissue architecture, and the extent of invasion. These assessments not only guide clinical decision-making but also serve as a basis for understanding tumor biology. However, while traditional histopathology has been invaluable, its reliance on subjective visual interpretation introduces variability and limits its capacity to harness the vast information contained in high-resolution histological images. The advent of computational methods has revolutionized this field, enabling the extraction of quantitative features from histopathological slides. These features range from measurements of nuclear size, texture, and cell density to spatial arrangement and tissue heterogeneity.

In our work, where we derived molecular subtypes of CRC based on gene expression profiles [9], we showed that the molecular subtypes correlate with tumour morphology – a histopathological variable which is not routinely assessed or reported. Most interestingly, tumours classified as molecular subtype D (20% of tumours) had the worse relapse-free survival and were characterized by increased expression of epithelial-mesenchymal transition (EMT) genes. The histopathological evaluation, however, led to the discovery that these tumours comprised often of more than 70% fibroblasts ("desmoplastic" morphotype). In consequence, this means that the observed high expression of the EMT genes is due to high fibroblast content and not to the stem-cell like (mesenchymal) tumour phenotype, as incorrectly interpreted in other studies. In addition, this important tumour subtype has escaped the attention of some cataloguing studies, such as The Cancer Genome Atlas (TCGA), which excluded tumours with tumour cell content lower than 80%. Molecular profiles thus must be interpreted with respect to histopathological evaluation. In addition, we observed multiple morphological patterns within the same tumour, and each can express a different molecular profile. A thorough histopathological evaluation of different tumour regions and micro-dissection of morphologically homogenous populations prior to molecular analyses was necessary for correct molecular classification. This, however, is in many studies impossible to achieve – in order to be correctly histopathologically evaluated, tumour specimens are routinely formalin fixed, and paraffin embedded (FFPE) after surgical excision to preserve histology. Most importantly, the examined specimen is usually taken from the invasive front of tumour on the colonic wall, since this region is the most important for characterization of tumour aggressiveness and its classification according to WHO standards. This part of tumour therefore cannot be stored as fresh frozen tissue, imposing important constraints on the methodology of sample collection. The studies performing molecular profiling from fresh frozen samples (considered of much better quality for molecular profiling) are therefore using material from a different tumour site, which can represent a different clonal population. This is often disregarded and introduces further bias into the interpretation of results.

In retrospect, it is natural to expect that tumor gene expression profiles represent a mixed signal derived from various cell types within the tumor microenvironment. Consequently, the observation that different tumor (cell) morphologies correlate with distinct molecular profiles is not surprising. This principle underlies the concept of deconvolution methods, which aim to estimate the proportions of different cell types present within tumor samples, providing a more nuanced interpretation of molecular data. However, while estimating the proportions of different cell types provides valuable insights, it overlooks a critical aspect of tumor architecture: the spatial organization of these cells within the tissue. Morphology, in contrast, inherently captures this spatial context and is relatively easy to assess in formalin-fixed paraffin-embedded (FFPE) samples. This is particularly true when enhanced by AI-driven image analysis software, which can standardize and automate morphological evaluations. Approaching tumor heterogeneity from a morphological perspective is not only more practical but also cost-effective, faster, and more universally applicable in clinical settings, as it leverages resources already available in most pathology departments.

In collaboration with Masaryk Memorial Cancer Institute, we collected multiple cohorts of colorectal samples, which enabled us to embark on a comprehensive exploration of the relationship between tumor morphology and molecular profiles, as well as the role of morphology in CRC heterogeneity.

Traditional digital image analysis in histopathology focuses primarily on the automatic extraction of predefined features. These typically include measurements such as nuclear size, cell density, the recognition and classification of different cell types, and their proportions within the tissue. These features are then quantified and statistically correlated with prognostic or diagnostic variables, offering an objective means of evaluating tumor morphology. While this approach has greatly improved the reproducibility of histopathological assessments, it remains constrained by human-defined criteria, effectively limiting the analysis to features that are already recognizable to the human eye.

In contrast, our approach diverged from this paradigm by leveraging molecular data to guide the extraction of image features, enabling us to capture patterns and relationships that go beyond what is visually discernible. By integrating histopathological images with gene expression profiles, we aimed to identify novel features reflective of underlying molecular mechanisms. This data-driven strategy opens possibilities for uncovering previously unrecognized biomarkers and relationships, providing a deeper understanding of tumor biology that is both more comprehensive and more closely aligned with the molecular heterogeneity of the disease.

Joint image and molecular analysis

First, we demonstrated how histopathological image features could be jointly analyzed with gene expression data, initially in the context of breast cancer [***12***], to identify meaningful correlations between morphology and molecular signatures. Histopathology images, while rich in information, are inherently complex, containing billions of pixels. To extract meaningful patterns, we employed a bag-of-features approach, which compresses the image data into essential patterns called codeblocks, identified using Gabor wavelets. This method enables the representation of each image as a histogram of codeblock frequencies, supplemented with extended features describing the spatial distribution of codeblocks, such as area, compactness, and skewness. This approach retained critical morphological information that is often overlooked in conventional analyses. The codebook was optimized through clustering to minimize overlap among tissue categories (e.g., fat, connective tissue, tumor nuclei), ensuring that the image representation was both sparse and discriminative. The resulting codeblocks captured three key morphological components: proliferation, invasion/differentiation, and isolated tumor nuclei (Figure 6).

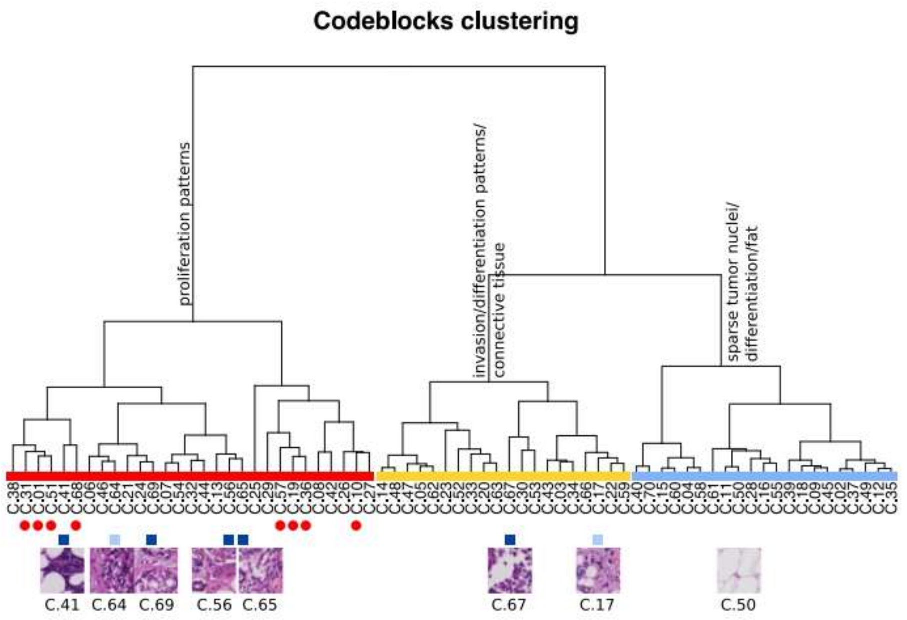


Figure 6. Hierarchical clustering of the codebook. Clustering the codeblocks led to identification of three major clusters, to which generic terms have been assigned. The codeblocks correlated with gene expression are marked with red dots. The codeblocks with potential prognostic value (in univariate analysis) are marked with blue squares (dark blue for p-value < 0.01, light blue for 0.01 ≤ p-value ≤ 0.05 (from [***12***])

The methodology also incorporated canonical correlation analysis (CCA) and other statistical tools to link these image features to gene expression, tumor size, grade, and relapse-free survival (RFS). A major contribution of this work was the development of an image-based prognostic score, derived from five key image features. This score was shown to be independent of genomic predictors and significantly improved prognostic models when combined with gene expression-based scores (Figure 7).

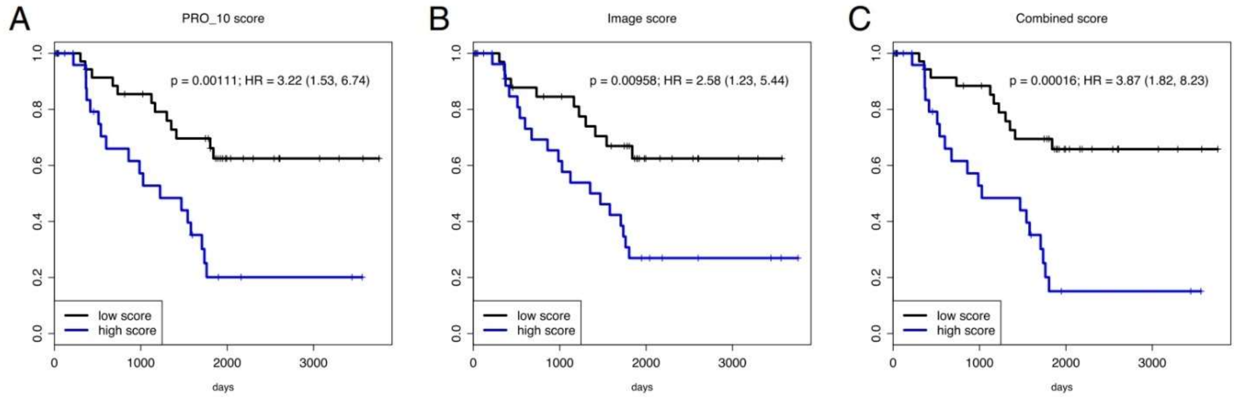
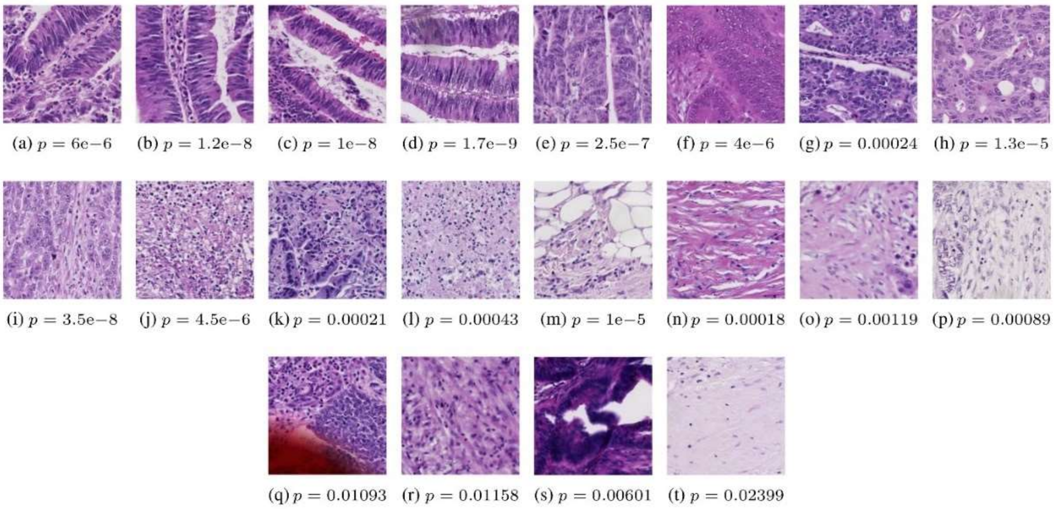


Figure 7. Kaplan-Meier curves for binarized scores. The genomic (a), image-based (b) and combined scores (c) were binarized by the respective median values into “low score” (low risk) and “high score” (high risk) categories. The combined score slightly improves on the genomic score (from [***12***]).

The code implementing this method was developed in R and made freely available for further research and application, laying the groundwork for broader integration of imaging and genomics in data mining and clinical practice [30].

Building on the observed link between tumor morphology and molecular profiles, we were the first to develop an image-based classifier capable of predicting CRC molecular subtypes from histopathological images [***13***]. Using histopathological slides from the PETACC-3 clinical trial, we analyzed a dataset of 300 tumor samples, which represented molecular subtypes A–E with the following frequencies: subtype A (n=21), B (n=140), C (n=37), D (n=81), and E (n=21). These samples were drawn from the PETACC-3 cohort of 458 molecularly annotated cases, focusing on those with high-quality images and sufficient tumor content, while excluding outliers and fragmented samples. The methodology involved processing hematoxylin-eosin (H&E) stained tumor sections, which were scanned at high magnification and subsequently downscaled to an equivalent magnification for computational efficiency. Tumoral regions were manually delineated based on expert pathologist annotations, ensuring that only relevant areas were analyzed. Local image features were extracted using a deep convolutional neural network (CNN) pre-trained on the ImageNet dataset, with the 4096-element descriptor vector from the penultimate layer reduced to 128 dimensions via principal component analysis (PCA). These local descriptors were pooled into global representations using Gaussian Mixture Models (GMMs), which facilitated the generation of a "visual codebook" summarizing key morphological features.



(a) $p = 6e{-}6$ (b) $p = 1.2e{-}8$ (c) $p = 1e{-}8$ (d) $p = 1.7e{-}9$ (e) $p = 2.5e{-}7$ (f) $p = 4e{-}6$ (g) $p = 0.00024$ (h) $p = 1.3e{-}5$

(i) $p = 3.5e{-}8$ (j) $p = 4.5e{-}6$ (k) $p = 0.00021$ (l) $p = 0.00043$ (m) $p = 1e{-}5$ (n) $p = 0.00018$ (o) $p = 0.00119$ (p) $p = 0.00089$

(q) $p = 0.01093$ (r) $p = 0.01158$ (s) $p = 0.00601$ (t) $p = 0.02399$

Figure 8. Top four prototypes associated with each subtype: (a–d) Subtype A, (e–h) Subtype B, (i–l) Subtype C, (m–p) Subtype D and (q–t) Subtype E. Under each image the corresponding *P* value from Wilcoxon rank-sum test is shown (from [***13***])

The image features were integrated into a multi-class **support vector machine (SVM)** classifier with a hierarchical design, optimized to first distinguish subtypes A and B from subtypes C, D, and E, before further separating individual subtypes. This decision tree structure was informed by misclassification patterns and the biological similarities between subtypes identified in previous studies. Model performance was assessed through 10-fold cross-validation, achieving an overall accuracy of approximately **85%**. Importantly, subtype-specific features, such as mucinous histology in subtype C or stromal desmoplasia in subtype D, were accurately captured by the classifier, reflecting the morphological heterogeneity across CRC molecular subtypes (Figure 8). Additionally, our analysis demonstrated that the image-based predictions could stratify patients by relapse-free survival (RFS) in a manner consistent with molecular subtyping, further validating the clinical relevance of the approach (Figure 9). While the SVM models we used for classification provided high accuracy, they are inherently difficult to interpret biologically. Future work aims to develop simplified models to facilitate biological interpretation, which is essential for clinical acceptance.

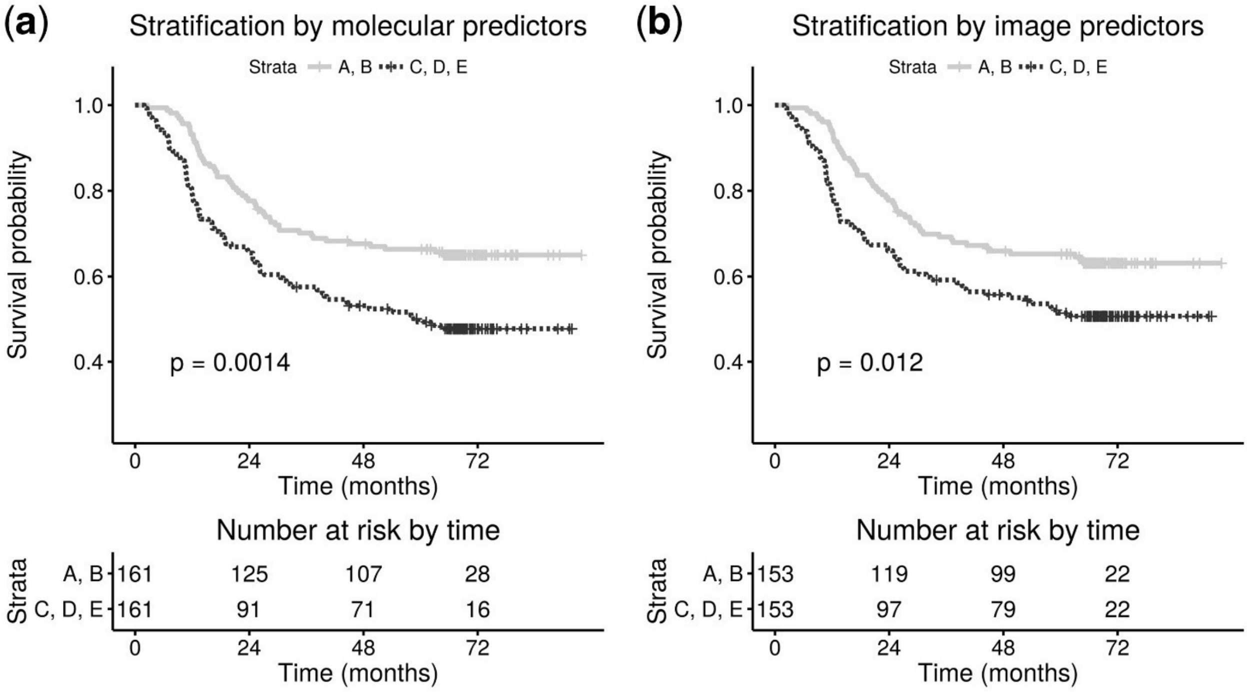


Figure 9. Survival analysis: risk of relapse stratified by (a) molecular subtypes and (b) image-based classifier. Subtypes A and B represent a lower risk group, while subtypes C, D and E a higher risk (from [13].)

Exploring molecular patterns of the morphotypes

One question, however, remained unanswered. We showed a clear association of our molecular subtypes with morphology, but the comprehensive molecular characterization of each morphological region was missing. In [*14*] we explored the transcriptomic landscape of the six morphotypes (CT, DE, MU, PP, SE and TB) and examined them alongside peritumoral regions, including normal adjacent tissue (NR) and supportive stroma (ST), to better understand how molecular programs map onto tumor histology. Using 111 unique primary CRC tumors across stages II (n=59), III (n=32), and IV (n=20), we macro-dissected 202 distinct regions, including 149 tumor regions, of which 126 were core samples containing at least 80% of a single morphological pattern (Figure 10). RNA extraction was performed on formalin-fixed paraffin-embedded (FFPE) histopathological slides, ensuring compatibility with archived clinical samples. Transcriptomic profiling was conducted using the Clariom D Array for human samples (Thermo Fisher Scientific), a platform that captures both coding and multiple forms of non-coding RNA.

The high-purity sampling allowed for a more accurate assessment of how distinct morphotypes contribute to the molecular heterogeneity of CRC. The samples originated from the COLOBIOME study, which we performed at Masaryk Memorial Cancer Institute in Brno [31].

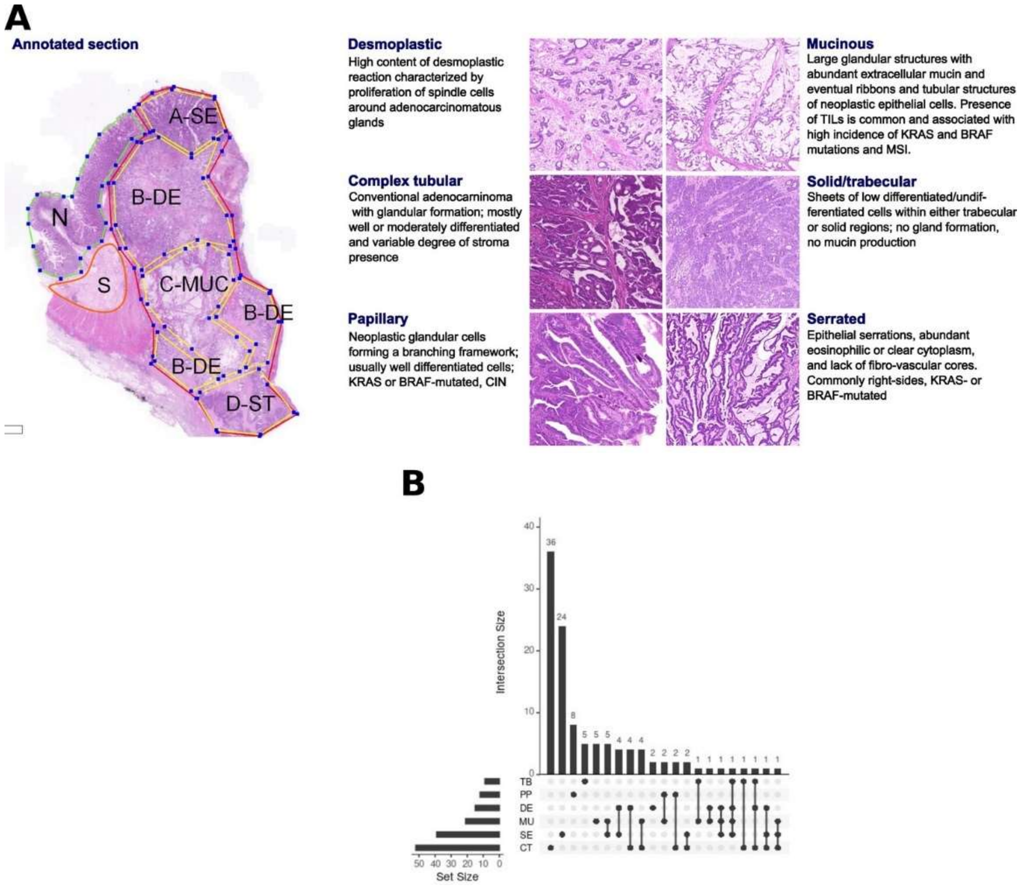

A
Annotated section
A-SE
B-DE
N
S
C-MUC
B-DE
B-DE
D-ST
Desmoplastic
High content of desmoplastic reaction characterized by proliferation of spindle cells around adenocarcinomatous glands
Mucinous
Large glandular structures with abundant extracellular mucin and eventual ribbons and tubular structures of neoplastic epithelial cells. Presence of TILs is common and associated with high incidence of KRAS and BRAF mutations and MSI.
Complex tubular
Conventional adenocaminoma with glandular formation; mostly well or moderately differentiated and variable degree of stroma presence
Solid/trabecular
Sheets of low differentiated/undifferentiated cells within either trabecular or solid regions; no gland formation, no mucin production
Papillary
Neoplastic glandular cells forming a branching framework; usually well differentiated cells; KRAS or BRAF-mutated, CIN
Serrated
Epithelial serrations, abundant eosinophilic or clear cytoplasm, and lack of fibro-vascular cores. Commonly right-sides, KRAS- or BRAF-mutated
B
Intersection Size
0
10
20
30
40
36
24
8
5
5
5
4
4
4
2
2
2
2
1
1
1
1
1
1
1
1
TB
PP
DE
MU
SE
CT
50
40
30
20
10
0
Set Size

Figure 10. Morphological patterns and their distribution in the dataset. (**A**) The six CRC morphological patterns of interest (morphotypes). Left: example of an original annotation used for macro-dissection and RNA extraction. Note that the original annotations in the image are not identical to the ones used in the main text. Here, A-SE stands for serrated (SE) in the text, B-DE for desmoplastic (DE) in the text, C-MUC for mucinous (MU) in the text, and D-ST for solid/trabecular (TB) in the text, respectively. Also, N indicates a tumor-adjacent normal epithelial region and S a supportive stroma region, respectively. Right: examples of morphotypes – complex tubular (CT), desmoplastic (DE), mucinous (MU), papillary (PP), serrated (SE), and solid/trabecular (TB). (**B**) Morphotype distribution per case (unique tumor) and intersections thereof: some cases had several morphotypes profiled (from [***14***]).

The morphotypes were characterized using gene set enrichment analyses (GSEA), and in silico deconvolution to identify differences in key biological processes, cell types, and pathway activity. Consistent with previous findings, MU and DE morphotypes (linked to CMS1 and CMS4) were enriched in fibroblast-associated signatures, TGF-β signaling, and immune response pathways, with DE further distinguished by inflammatory CAFs (IL-iCAF). Conversely, epithelial-rich SE and PP morphotypes showed downregulated EMT and KRAS signaling but upregulated MYC target pathways, reflecting their connection to the serrated oncogenic pathway. Interestingly, the CT and TB morphotypes demonstrated active proliferation and basal cell signatures, with TB also sharing stromal characteristics such as active TGF-β signaling with MU and DE (Figure 11).

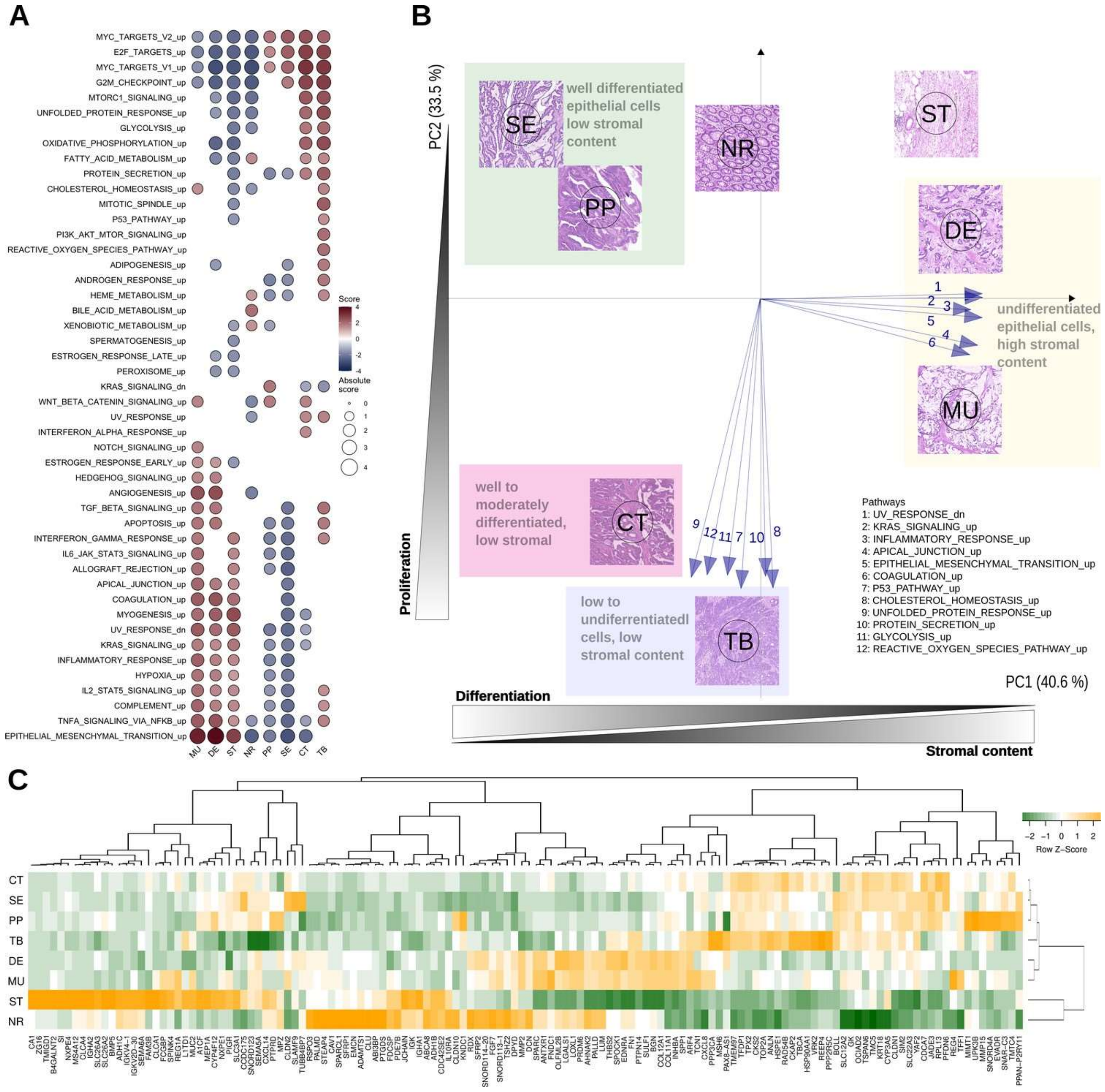


Figure 11. Top differentially expressed genes and hallmark pathways. (A) GSEA scores for hallmark pathways in the six morphotypes and two non-tumoral regions. Only pathways with statistically significant scores are shown. (B) Principal component analysis of hallmark pathways: the median profiles of the six morphotypes (CT: complex tubular, DE: desmoplastic, MU: mucinous, PP: papillary, SE: serrated, and TB: solid/trabecular) and the two non-tumoral regions (NR: tumor-adjacent normal and ST: supportive stroma) are projected onto the space defined by first two principal components (74% of the total variance). The top pathways contributing to the principal axes are shown as well. See also Figure 3—figure supplement 1. (C) Heatmap of top 5 up- and down-regulated genes for each of the six morphotypes (from [*14*]).

An important aspect of this study was the exploration of intra-tumoral heterogeneity. By analyzing matched regions within the same tumor, we showed that molecular classifiers like CMS are less stable at the regional level, with 60% of tumors displaying discordance between CMS assignments in whole-tumor profiles versus individual regions (Figure 12).

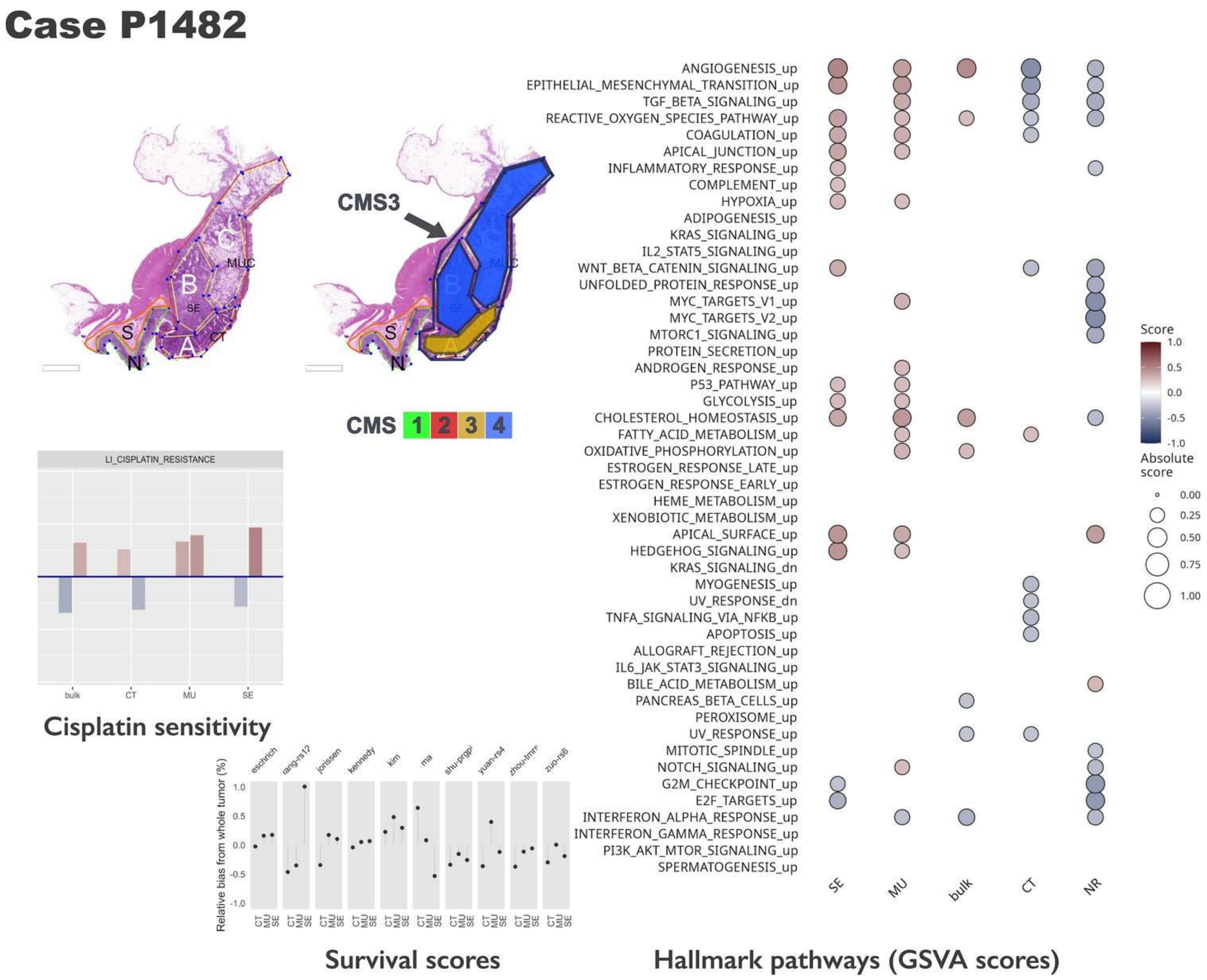


Figure 12. Intra-tumoral heterogeneity case study. For the same case, different CMS labels are assigned to regions and whole tumor profile. The hallmark pathways show various levels of activation (as computed by GSVA) within the same section. The relative change in prognostic scores indicates potential underestimation of risk for some signatures, while others appear to be stable across tumor. Note that in the pathology section image, the original annotations were preserved, and they are not identical to the ones used in the main text. Here, MUC stands for mucinous (MU) in the text. Also, N indicates a tumor-adjacent normal epithelial region and S a supportive stroma region, respectively (from [*14*]).

Prognostic gene expression signatures also varied significantly across regions, with some morphotypes showing scores more than 50% higher than the corresponding whole-tumor score. This suggests that whole-tumor profiling may underestimate the risk in morphologically heterogeneous tumors. Our findings showed the need to account for tumor morphology in molecular profiling studies and that anchoring expression profiles to histopathological morphotypes can serve as a practical alternative to spatial transcriptomics, which remains challenging to implement in routine practice. To support further research, we developed a web application [32] for interrogating gene expression profiles in various morphological regions, providing a valuable resource for the broader scientific community.

How heterogenous the tumours are in terms of morphologies?

Given these observations, we naturally turned our attention to exploring the full extent of tumor morphological heterogeneity across CRC cases and its potential clinical implications. In our recent study [1], we aimed to address this question by combining traditional pathology with cutting-edge AI-driven image analysis. Specifically, we sought to quantify the diversity of tumor morphotypes within individual cases and to assess the clinical relevance of this heterogeneity.

We began with a pilot analysis of 22 CRC tumors, sampling four histological sections per tumor (n=88 slides) and employing three expert pathologists to evaluate the dominant, secondary, and tertiary morphologies in each section. This initial assessment revealed a high degree of morphological heterogeneity, with most tumors exhibiting 2–3 dominant morphotypes across different sections. The complex tubular (CT) morphotype was the most frequently observed, while desmoplastic (DE) the least observed. Inter-pathologist variability was minimal for CT and more prominent for DE and MU (mucinous) morphologies, emphasizing the need for a standardized, objective method to classify these patterns.

To scale up the analysis, we developed an AI-based deep learning model trained on the annotations from the pathologists. This model was applied to an expanded cohort of 161 CRC cases (n=644 slides), allowing us to systematically characterize the distribution of morphotypes and their combinations. The AI-guided analysis confirmed the findings of the pilot study, with over 50% of tumors exhibiting more than two dominant morphotypes and medium to high morphological diversity, as measured by a normalized Shannon index (Figure 13).

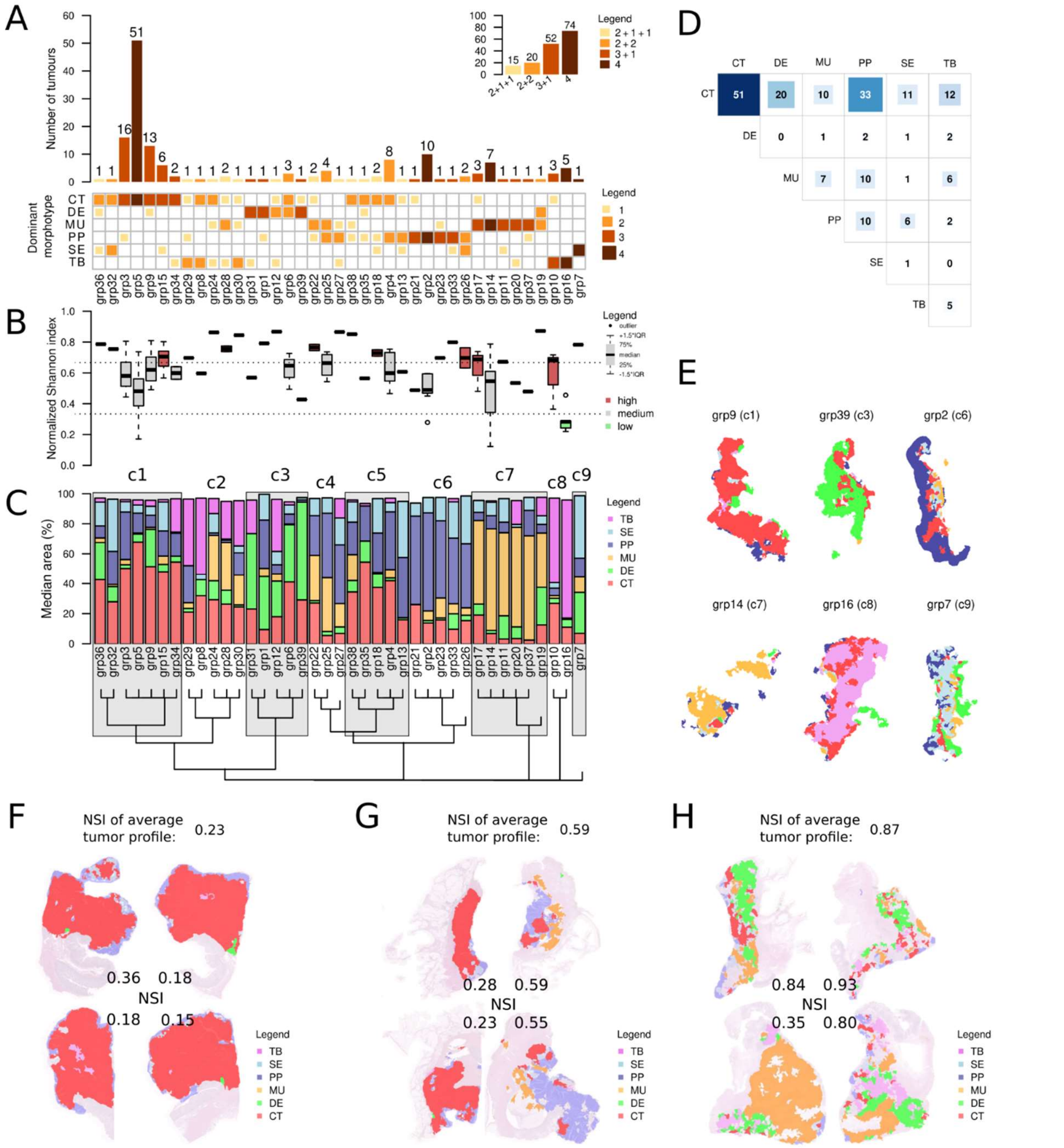


Figure 13. **A.** Observed intratumoral patterns of dominant morphotype combinations (IPDMCs) and their frequency (main barplot) and frequency of their distribution patterns (embedded top right barplot). **B.** Distribution of normalized Shannon index of median tumor profiles in the IPDMCs. **C.** Median morphotype area in the IPDMCs and their further clustering into 9 clusters. **D.** Frequencies of pairwise combinations of dominant morphotypes in the sections. **E.** Examples of representative tumor morphological areas of slides from selected IPDMCs clusters as identified by image analysis. **F-H.** Examples of intratumoral morphological heterogeneity as assigned by image analysis over four examined slides/blocks. Values of normalized Shannon index (NSI) of each slide and the average tumor profile are shown. **F.** Tumor with low heterogeneity across all slides, expressing one dominant morphotype (CT). **G.** Tumor with low heterogeneity in two slides and medium heterogeneity in two slides, expressing two dominant morphotypes (CT and PP) **H.** Tumor with high heterogeneity in all four sections, expressing two dominant morphotypes (DE and MU) (from [1]).

Importantly, this diversity was not itself associated with clinical variables, but the proportion of specific morphotypes—such as DE and MU—correlated strongly with outcomes. For example, tumors with higher proportions of DE morphotype were associated with advanced T-stage, N-stage, metastasis, and shorter relapse-free survival (RFS), while MU was linked to MSI, right-sided location, and poorer overall survival (OS).

These findings highlighted the critical need to consider intratumoral heterogeneity when performing molecular analyses. For instance, we observed that morphotypes such as MU and PP often coexisted with other morphologies, potentially reflecting shared oncogenic programs like *KRAS* or *BRAF* mutations. Conversely, the SE morphotype, which is associated with the serrated neoplasia pathway, was rarely found alongside MU. Similarly, CT, the most common morphotype, often combined with other morphologies, while TB—a hallmark of dedifferentiation—was predominantly found in tumors with low diversity and was largely independent of specific oncogenic pathways.

3.5. TUMOR MICROENVIRONMENT AND MICROBIOME

At the same time, and somehow in parallel, more and more attention is paid to CRC tumour microenvironment and even more importantly to the role of gut microbiota. Human gut microbiota is composed of four major domains of life of which vast majority are Bacteria (95%), the rest being Archaea, Eucaryota and Viruses. Gut microbiota outnumbers 10 times the number of human cells in the human body and comprises majority of mammalian-associated microbes. It is referred to as commensal intestinal microbiota and forms a versatile microecosystem that changes its composition in response to the host’s development, diet, or disease state [33]. Most dense and metabolically active microbial community resides in the large intestine, comprised mainly of anaerobic bacteria of two phyla: Firmicutes and Bacteroidetes, accompanied by Actinobacteria, Proteobacteria and Verrucomicrobia. In a healthy organism, the gut microbiota is in a symbiotic relation with the human host and contributes to controlling the intestinal epithelial homeostasis, to food digestion, to synthesis of certain vitamins and to defense against pathogens. This has a great impact on the set-up of the human immune system, which uses specific mechanisms for discrimination of between harmful and helpful microbial species: immune exclusion (secretory IgA antibodies present in mucosa layer selectively block antigens and pathogenic microbiota from accessing cell epithelial receptors) and immunosuppression (recognizing antigens of pathogenic and commensal bacteria via Toll-like receptors - TLRs) [34,35]. The tumour microenvironment contains several different immune cell types, including tissue-associated macrophages (TAMs) and other innate immune cells, as well as T cells and B cells, which communicate with each other and the other cells in the tumour microenvironment via direct contact or via cytokine and/or chemokine signalling to control tumour growth. TAMs primarily promote tumour growth, and high numbers of TAMs generally correlate with cancer progression.

Dysbiosis – chronic alteration of gut microbiota – is reported in many diseases, such as autoimmune diseases or even colon cancer and there is growing evidence that development of these diseases is influenced by microbiota - human immune response interactions [33]. Recent studies show that bacteria adherent to colorectal adenomas or carcinomas are different from bacteria adherent to healthy mucosa [36]. This is a result of changes in the local tumour microenvironment, which has decreased pH and changed nutritional conditions as a consequence of altered metabolism of tumour due to hypoxia [37]. Bacteria can promote colon cancer development or change the tumour invasion potential through immunomodulation

[38,39] or metabolic activity – through production of specific toxins inducing DNA damage responses [40]. This is enhanced by defects in barrier function of the gut, which allow luminal bacteria to translocate to epithelial layer and directly influence host cells. Overall, the evidence of microbiome importance in colon cancer development is so overwhelming that a bacterial driver-passenger model for colorectal cancer development and progression was suggested [36] as an alternative to the broadly accepted driver-passenger mutational adenoma-carcinoma model.

Our expertise in the microbiome and its role in CRC was summarized in a review article published in *Klinická onkologie* [***15***]. This journal, written in Czech and targeted at clinical oncologists, aimed to bridge the gap between basic research and clinical practice by providing a comprehensive overview of the microbiome's role in CRC development, progression, and potential therapeutic implications. This work served as a foundation for the more detailed experimental studies that followed. In addition, we contributed a chapter titled *Mikrobiom v solidních nádorech* to the book *Mikrobiom a zdraví* [41]. This chapter explored the microbiome's interactions with solid tumors in depth, including microbiota residing on the tumor surface, within the tumor, and even inside tumor cells. It also examined the mechanisms of microbiome-tumor interaction, such as immune modulation and metabolic influence, and discussed the potential for directly targeting tumors through microbiota-based therapies. Together, these contributions reflect our multifaceted approach to understanding and leveraging microbiome in cancer research and clinical application.

Gut microbiota plays an important role also in cancer therapy. Microbiota influences drug metabolism, immune responses, and the tumor microenvironment, thereby impacting the effectiveness of chemotherapy, immunotherapy, and radiotherapy. For example, *Akkermansia muciniphila* has been shown to enhance the efficacy of immune checkpoint inhibitors by stimulating anti-tumor immune responses [42]. In contrast, antibiotic-induced microbiome depletion can impair therapeutic efficacy, as observed in studies linking antibiotic use with reduced responses to both immunotherapy and chemotherapy [43]. Furthermore, certain microbes can either promote or inhibit tumor growth through their effects on drug metabolism, such as the modulation of gemcitabine efficacy by *Gammaproteobacteria* [44]. Microbiome can be perceived as both a therapeutic ally and a potential barrier, and it is of crucial importance to consider microbiota modulation as a complementary approach to optimize cancer treatments.

It has been long recognized, that bacteria are capable of penetrating and moving within the tumour [45], making them a perfect candidate for anticancer agents. When using bacteria for treatment, tumour regression can be achieved by native bacterial cytotoxicity caused by sensitization of the immune system and competition for nutrients [46]. This effect, however, can be hampered by the immune system preventing intra-tumoural bacterial dissemination [47]. One hypothesis we put forward is that the intra-tumoural presence of commensal bacteria (either recruited by the tumour or opportunistic) helps the tumour escape the immune system, since these bacteria are not recognized as pathogens. If this hypothesis can be validated, it could serve as the basis of a bacterial-targeted treatment.

It is our strong belief that the identification of gut microbiota specific to treatment-resistant tumours is a key step towards a finer patient population stratification and more targeted therapies.

The need for multimodal approach

Molecular profiling, however powerful, constitutes only one modality of exploration of the complex picture of CRC heterogeneity. The machinery of molecular events adapts swiftly to the signals from its surrounding microenvironment, which plays an important role in shaping the tumour phenotype. Tumour and patient metabolome profiling is currently in its renaissance and is being exploited for identification of marker metabolites defined as surrogate indicators of colorectal cancer development [48]. Differences in metabolic profiles were found not only between normal and cancer tissue, but also within subtypes of CRC [49]. The metabolic and inflammatory milieu within the tumour microenvironment may affect the function and phenotype of tumour cells, irrespective of genotype.

While we are witnessing increased interest in characterizing the gut microbiome from cancer clinical perspective, this research applied to colorectal cancer lags behind tumour molecular profiling by several years. Some studies tried to incorporate information on tumour associated microbiome in order to improve the accuracy of the existing patient CRC prognostic score [50] or developed new screening/prognostic models [51]. However, despite consistent patterns of gut microbial disruption in comparison to healthy individuals [52,53], the variability between diseased individuals remains too high. One source of this variability is the type of diet. Another source, however, can again be assigned to tumour molecular heterogeneity and the respective tumour metabolic profile, which might influence the tumour microbiota. Due to this, we suggested, that any study aiming at unveiling the role of microbiota in colorectal cancer progression or response to therapy should investigate the presence and distribution of bacteria and immune cell types accounting for the intra-tumoural heterogeneity and metabolism. If microbiota is to answer some of the key outstanding questions about CRC heterogeneity that are not explained by molecular profiling, we have to move from simple healthy-tissue/adenoma/carcinoma correlation studies towards complex multimodal approaches. A new approach is definitely needed, that is data-driven and can cleverly and efficiently mine and combine all the modalities (molecular data, clinical data, histopathology, metabolism and microbiome) and not only catalogue the existing correlations but also generate sound hypotheses that can be tested in further functional analyses.

This perspective motivated us to submit the AZV research project (COLOBIOME), which aimed to integrate microbiome and tumour microenvironment analyses into the study of colorectal cancer heterogeneity. Through this project, we successfully established a prospective cohort of approximately 200 stage I–IV CRC patients. This cohort includes an extensive array of samples: stool samples, tumour and adjacent visually normal mucosa swabs for microbiome profiling, tumor resections preserved as FFPE and fresh-frozen tissues, as well as peripheral blood collected at the time of diagnosis.

Methodological considerations of microbiome studies in clinical samples

Studying the microbiome in clinical samples is inherently challenging due to numerous technical and biological factors that can influence the quality and reproducibility of results. In our study [***16***], we systematically evaluated the impact of stool sampling methods and DNA isolation kits on quality of extracted DNA and estimation of bacterial composition and diversity using 16S rRNA sequencing on the MiSeq Illumina platform. Thanks to this study, we gained insights into the methodological aspects that need to be standardized to ensure robust microbiome profiling in our further studies.

Sixteen volunteers provided samples from a single stool using three sampling kits: stool container (SK1), flocked swab (SK2), and cotton swab (SK3). DNA was extracted using two isolation kits, PowerLyzer PowerSoil (PS) and QIAamp DNA Stool Mini Kit (QS), resulting in 96 samples for analysis. User preference evaluations showed that 100% of participants favored the stool container (SK1) for ease of use, while 81.25% found the flocked swab (SK2) least convenient due to challenging handling. DNA quality assessments revealed higher yields with the QS kit, but PS preserved better DNA integrity, particularly when paired with stool containers. Interestingly, stool container samples also exhibited reduced PCR inhibitors, enhancing downstream processing efficiency. While bacterial diversity metrics (Chao 1 and

OTUs) were influenced by both sampling and isolation methods, PS consistently extracted more Gram-positive bacterial taxa due to its robust bead-beating procedure.

Bacterial composition analysis confirmed that both the sampling and isolation methods significantly influenced taxonomic abundance, particularly at higher taxonomic levels (phylum, class, order) (Figure 14). PS exhibited greater efficiency in recovering Gram-positive taxa, a trend attributed to its superior cell lysis capabilities. Notably, stool container samples resulted in higher bacterial diversity, likely due to optimized sample dilution during preprocessing.

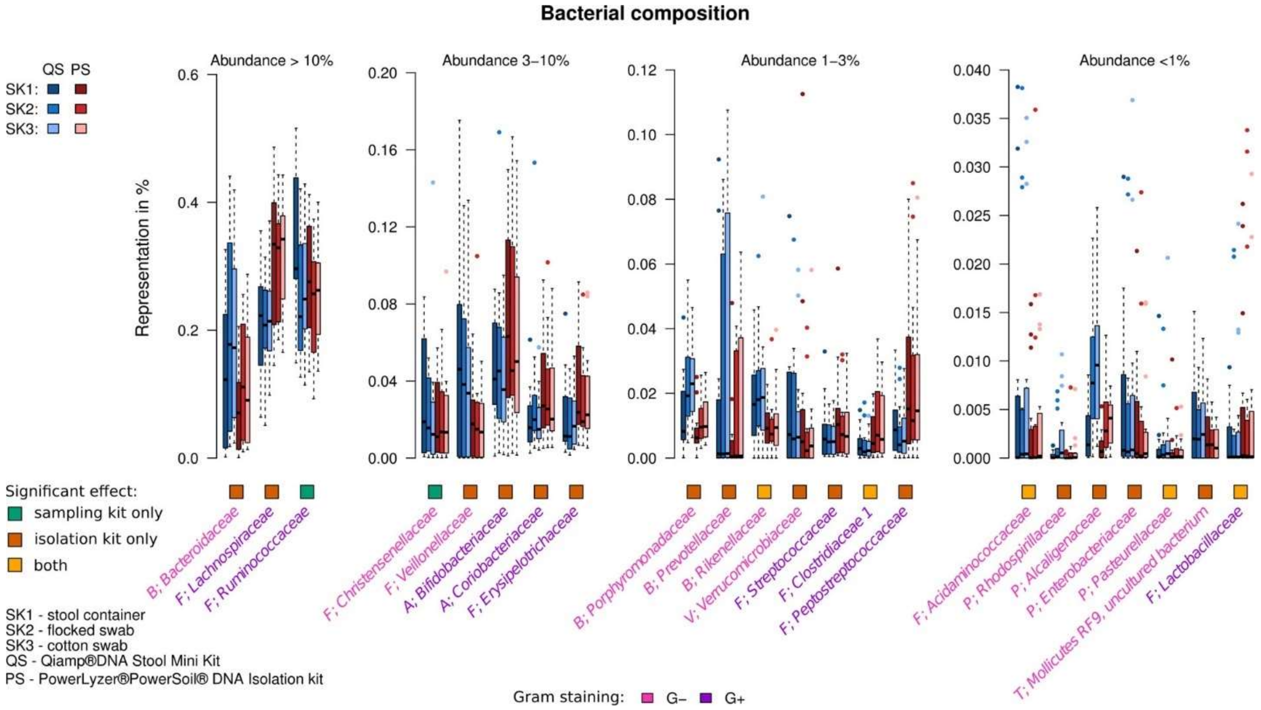


Figure 14 Distributions of relative abundances of significantly affected taxa at family level. Four graphs represent families divided according to third quartile of their abundance. Only taxa that passed the filtering criteria (maximum abundance >1%), significantly affected by isolation or sampling kit are shown. The colored squares below the graph indicate whether the family was affected significantly by the sampling kit only, the isolation kit only or both. (from [*16*])

The tumour mucosa microbial subtypes

The microbial composition within the colorectal cancer (CRC) tumor microenvironment represents a pivotal factor in understanding tumor biology and its progression. In our study [*17*], we adopted a microbial community-centric approach to comprehensively characterize the heterogeneity of the tumor-associated microbiome across three distinct sampling environments: tumor mucosa, adjacent visually normal mucosa, and stool samples. Utilizing a cohort of 178 CRC patients (stages 0–IV) from the COLOBIOME project, and analyzing 483 samples, our aim was to explore the microbial landscape and its association with clinical variables, while addressing limitations of earlier studies that were either species-centric or underpowered in sample size. By focusing on microbial communities rather than individual species, we provided a different perspective into the tumour microbial heterogeneity. Our final result was identification of CRC tumor mucosal microbial subtypes.

To ensure robust methodology, 16S rRNA sequencing was used for microbial profiling, with data processing performed using state-of-the-art compositional data techniques. Prior to analysis, zero multiplicative replacement and centered log-ratio (clr) transformations were applied to address the compositional nature of microbiome data. Microbial diversity was evaluated using alpha diversity metrics, such as Chao 1 and observed species, while beta diversity was analyzed based on Aitchison distance matrices. Co-occurrence patterns among microbial taxa as well as clusters of tumours with similar microbial compositions were identified through hierarchical clustering.

To identify differences in diversity and bacterial composition across environments and their associations with clinical variables, we applied comprehensive statistical analyses, including the Friedman test, rank regression, and permutational multivariate analysis of variance (PERMANOVA). Multiple testing corrections were conducted using the Benjamini-Hochberg procedure, with a false discovery rate (FDR) threshold set at < 0.1.

The analysis of the tumor microbiome associations with clinical variables revealed numerous associations. Notably, higher-grade tumors (grade 3) were characterized by an increased abundance of the potentially pathogenic genera such as *Fusobacterium*, *Campylobacter*, *Leptotrichia*, *Selenomonas*, and *Prevotella* in tumor mucosa, reflecting their potential role in tumor progression and aggressiveness. These associations were particularly pronounced in right-sided tumors, where high-grade tumors were enriched in genera such as *Prevotella* and *Selenomonas*. In contrast, lower-grade tumors (grade 1 and 2) and left-sided tumors exhibited a depletion of pathogenic genera and an enrichment of commensal species such as *Bifidobacterium*, *Ruminococcaceae UCG-010*, and *Victivallis*. Tumor location was also a critical determinant of microbiome composition, with distinct microbial signatures observed for right-sided and left-sided tumors, reflecting the well-known biological differences between these tumor types. Advanced tumor stages (pT3/pT4) were associated with increased abundance of genera such as *Peptoclostridium* and *Parvimonas* in tumor mucosa, while metastasis status primarily influenced the stool microbiome, with genera like *Akkermansia* enriched in patients with local or distant metastases.

To characterize the microbial heterogeneity of tumor mucosa while excluding potential stool contaminants, we focused solely on species that were statistically significantly more abundant in tumor mucosa compared to stool. Overall, 121 genera showed significant differences in abundance across sample environments, leading to the definition of five microbial categories: *tumor genera* (enriched in tumors compared to stool), divided further into *mucosa genera* (shared enrichment in tumor and visually normal mucosa) and *tumor-specific genera* (enriched only in tumor mucosa). Then, we defined *stool genera* (enriched in stool), and no-difference genera (consistent abundance across sample types). The analysis uncovered 57 genera enriched in tumor mucosa compared to stool, 16 of which were defined as tumor-specific genera, uniquely associated with tumor tissue and absent in adjacent normal mucosa. Notably, these tumor-specific genera predominantly consisted of genera comprising oral pathogens such as *Fusobacterium, Parvimonas, Campylobacter, and Leptotrichia,* supporting their potential role in tumorigenesis. Similarly, bacterial groups dominated by gut commensals, such as *Ruminococcus* and *Bacteroides*, were primarily found in stool samples, emphasizing the distinct microbial ecosystems between mucosal and luminal compartments.

The bacteria were classified into six groups, labeled B1–B6. Groups B1 and B2 predominantly represented typical gut microbiome members. The B1 group included the five most common and abundant genera: *Fusobacterium*, *Lachnoclostridium*, *Bacteroides*, *Escherichia-Shigella*, and an uncultured genus from the *Lachnospiraceae* family. Nearly all tumors contained at least three of these genera, with 78.7% containing all five. These bacteria exhibited high co-occurrence across sample types, except for *Fusobacterium*, which was primarily found in mucosal samples. Group B4, referred to as the *Selenomonas* group, was exclusively composed of oral microbiome genera, enriched in *Selenomonas*. Groups B3 and B5 also primarily consisted of oral microbiome genera, which exhibited significantly lower incidence in stool samples, being absent in 45.7%–94.1% of cases where they were present in tumor mucosa. Finally, Group B6 comprised 27 less common species with incidence rates ranging from 0% to 37% (median 11.1%).

The new tumour microbial classification was proposed on the 57 *tumour genera* and comprised of three tumor microbial subtypes (TMS1–TMS3), each associated with distinct clinical and microbial features (Figure 15).

TMS1: High Pathogen Burden and Biofilm Association

TMS1, representing 26% of tumors, was characterized by the highest microbial pathogen burden, with a high abundance of potential oral pathogens and bacteria associated with advanced tumor progression. This subtype contained all microbial groups (B1–B4) and was enriched in genera such as *Fusobacterium*, *Campylobacter*, *Leptotrichia*, *Peptoclostridium*, and *Selenomonas*. These bacteria have been linked to biofilm formation and cancer-associated inflammation, which might explain the more aggressive features of TMS1 tumors. Clinically, TMS1 was associated with right-sided tumors (60.9%), higher-grade tumors (58.7% grade 3), advanced pathological stages (95.6% pT3/pT4), and a higher prevalence of microsatellite instability-high (MSI-H, 34.8%) and BRAF mutations (15.2%). This subtype likely represents a biologically distinct group of tumors enriched in microbial biofilms, which could drive local inflammation and promote tumor progression.

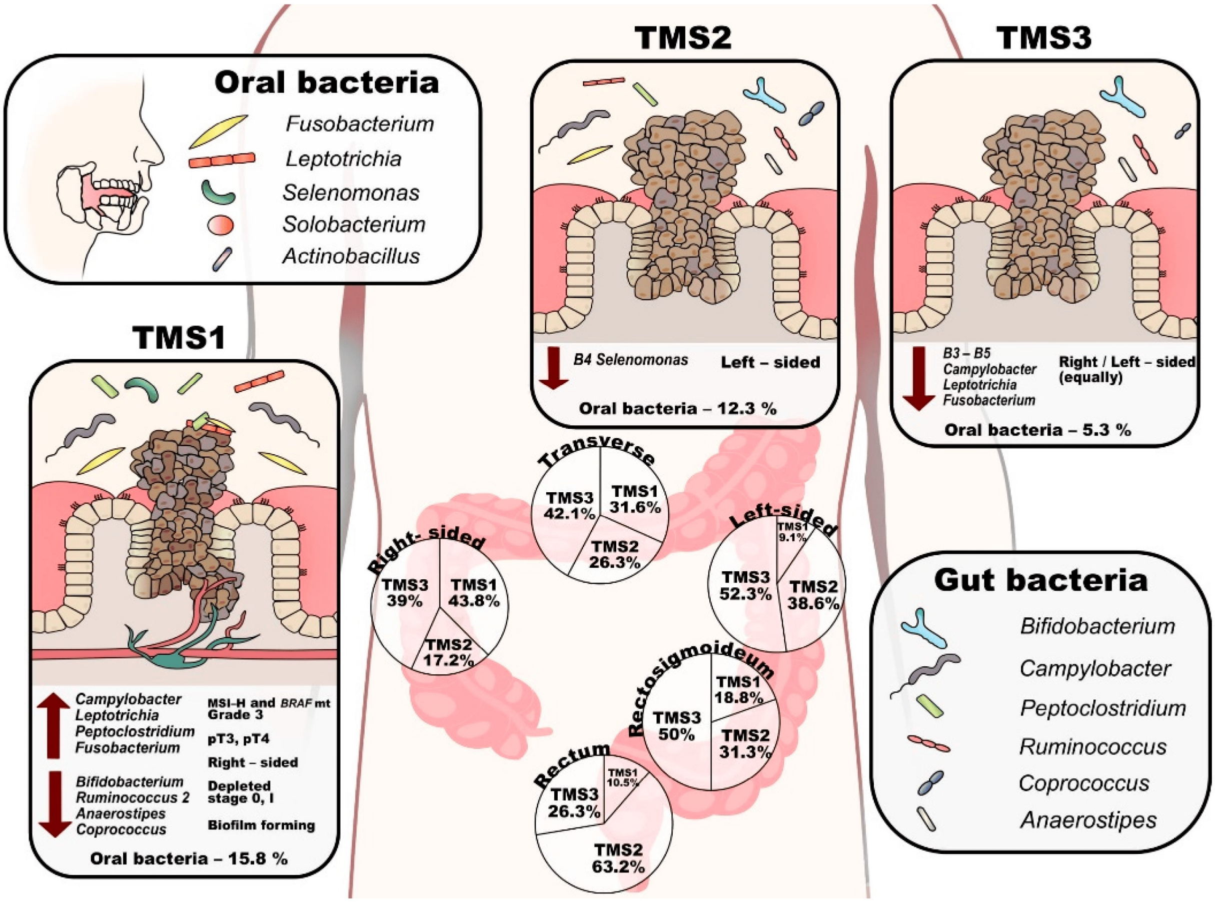


Figure 15. The characteristics of the three tumour microbial subtypes. (TMS—tumour microbial subtypes, pT—tumour pathologic stage, MSI-H— microsatellite instability-high) (from [*17*])

TMS2: Intermediate Pathogen Burden and Left-Sided Tumors

TMS2 accounted for 31% of tumors and displayed an intermediate pathogen burden. Unlike TMS1, this subtype lacked bacteria from the *Selenomonas* group (B4) but included other oral and gut-associated genera such as *Leptotrichia*, *Granulicatella*, *Aggregatibacter*, and *Veillonella*. TMS2 tumors were predominantly left-sided, including rectosigmoid and rectal tumors (70.9%), and exhibited a more heterogeneous microbial composition. This subtype could be further divided into two groups: TMS2a, enriched in oral bacteria such as *Neisseria* and *Granulicatella*, and TMS2b, enriched in gut-associated bacteria such as *Tyzzerella 4* and *Hungatella*. TMS2 mucosal microbiomes also showed a higher abundance of commensal bacteria such as *Haemophilus* and *Sutterella*, indicating a more balanced microbial environment compared to TMS1.

TMS3: Low Pathogen Burden and Commensal-Enriched Microbiome

TMS3, the largest subtype, comprised 43% of tumors and was defined by a low microbial pathogen burden. This subtype had a reduced presence of oral pathogens and biofilm-associated bacteria and was characterized by a higher proportion of commensal gut bacteria. TMS3 tumors were equally distributed between right-sided and left-sided locations, with a notable enrichment of lower-grade tumors (15.6% grade 1). TMS3 was further divided into two subgroups: TMS3a, associated with an increased presence of *Incertae Sedis* from the *Erysipelotrichaceae* family and *Tyzzerella 4*, and TMS3b, characterized by genera such as *Clostridium sensu stricto 1*, *Ruminococcaceae UCG-013*, and *Lachnospiraceae Incertae Sedis*. Interestingly, TMS3 included all tumors that lacked *Fusobacterium* in both tumor mucosa and stool samples, suggesting a distinct microbial profile compared to the other subtypes.

In conclusion, this study extended the characterization of the colorectal cancer microbiome in several important directions. By analyzing 483 samples from 178 patients, we identified bacterial genera previously unassociated with colorectal cancer mucosa or clinical variables, revealing novel avenues for understanding their roles in disease progression. Our focus on microbial community-level analysis rather than species-centric approaches allowed us to describe three major tumor-microbial subtypes, each differing in microbial composition, associations with clinical parameters, and what we define as microbial pathogen burden—highlighting their potential relevance to tumor aggressiveness and progression.

The complementary nature of the sampled environments—tumor mucosa, visually normal mucosa, and stool—provided insights into the distinct contributions of the microbiota across these niches. While tumor mucosa and visually normal mucosa reflected tumor-specific variables, such as grade and location, stool microbiomes were more influenced by the presence of metastases and overall disease progression. It is evident that combining both mucosal and stool sampling is essential for gaining a more comprehensive understanding of CRC microbiome dynamics.

Although this study provided valuable insights, the absence of validation data and the potential influence of dietary and lifestyle factors limit broader applicability. Further investigations with larger, geographically diverse cohorts are necessary to confirm and refine these findings. Nonetheless, the ability to associate tumor microbial subtypes with clinical variables suggests the potential for leveraging the microbiome in CRC management. Tailored strategies, such as diet modifications, probiotics, or antimicrobial interventions, may emerge as valuable additions to current therapeutic approaches. This study represented a significant step forward, offering new perspectives for exploring microbiome-related treatments and biomarkers in colorectal cancer.

Our unique CRC dataset provided the foundation for our involvement in the H2020 project ONCOBIOME [54], which enabled us to further enhance our tumor sample data by incorporating whole metagenome sequencing (WMGS) and stool-derived miRNA profiling.

3.6. OMICS BIOMARKERS IN DIAGNOSTICS AND THERAPY OF CRC

mRNA Biomarkers for Predicting FOLFIRI Treatment Response

Personalized treatment strategies in colon cancer remain a major clinical need, particularly in optimizing adjuvant chemotherapy selection for Stage III patients. In our study using the PETACC-3 clinical trial cohort, we evaluated the predictive value of ABCG2 and topoisomerase 1 (TOP1) mRNA expression for assessing the benefit of irinotecan-based therapy (FOLFIRI) [***18***]. We analyzed mRNA expression data from 580 Stage III colon cancer patients randomized to receive either 5-fluorouracil/leucovorin (5FUL) alone or in combination with irinotecan (FOLFIRI). Patients were stratified into two biomarker-defined groups: a "resistant" group characterized by high ABCG2 and low TOP1 expression (n = 216) and a "sensitive" group encompassing all other expression profiles (n = 364).

Applying Cox proportional hazards regression, Kaplan-Meier survival analysis, and log-rank testing, we demonstrated that patients classified as "sensitive" derived significant benefit from FOLFIRI, with improved recurrence-free survival (HR: 0.63, p = 0.016) and overall survival (HR: 0.60, p = 0.02) compared to the "resistant" group. These associations were even stronger in microsatellite-stable (MSS) and microsatellite-instability-low (MSI-L) patients (n = 470), while no survival differences were observed when patients received 5FUL alone. This suggested that the ABCG2/TOP1 mRNA profile may serve as a clinically relevant biomarker for predicting responsiveness to irinotecan-based chemotherapy.

Fecal microRNA Signatures for Non-Invasive CRC Diagnosis

Current colorectal cancer (CRC) screening programs rely on fecal tests, which often lack the sensitivity needed to detect early-stage tumors and precancerous lesions. Thanks to our participation in the H2020 project ONCOBIOME, I had the opportunity to address this limitation and explore stool microRNA (miRNA) profiles as potential biomarkers for non-invasive CRC detection, leveraging a comprehensive multi-cohort study and explainable machine-learning approaches.

In this study [*19*] conducted 1,273 small RNA sequencing experiments across multiple biospecimens, analyzing fecal samples from both an Italian and a Czech cohort (155 CRC patients, 87 adenomas, 96 individuals with other intestinal diseases, and 141 colonoscopy-negative controls) (Figure 16).

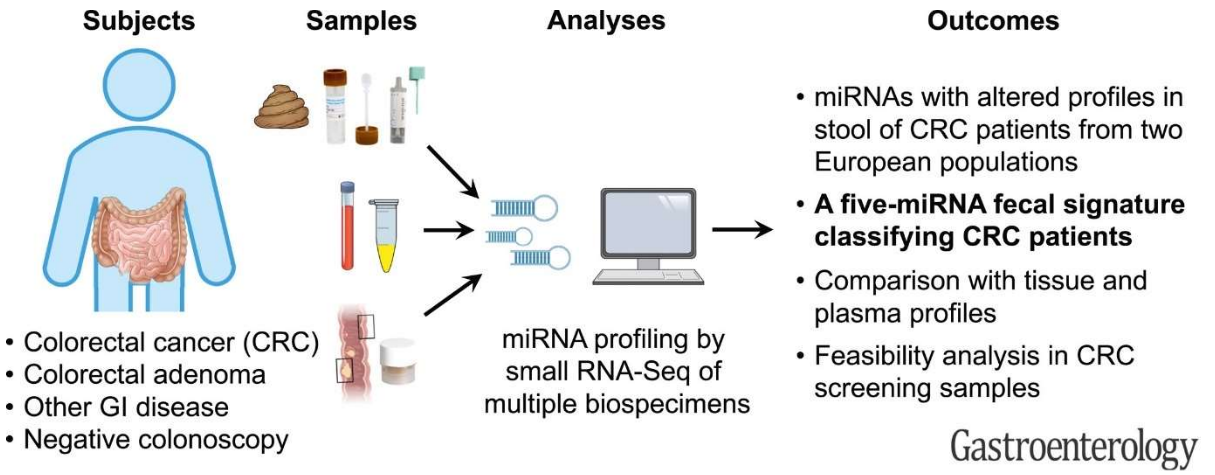

Subjects
Samples
Analyses
Outcomes
• Colorectal cancer (CRC)
• Colorectal adenoma
• Other GI disease
• Negative colonoscopy
miRNA profiling by small RNA-Seq of multiple biospecimens
• miRNAs with altered profiles in stool of CRC patients from two European populations
• A five-miRNA fecal signature classifying CRC patients
• Comparison with tissue and plasma profiles
• Feasibility analysis in CRC screening samples
Gastroenterology

Figure 16. Graphical abstract of the miRNA biomarker study (from [*19*]).

Through systematic analysis, we identified a robust 5-miRNA signature (miR-149-3p, miR-607-5p, miR-1246, miR-4488, and miR-6777-5p) capable of distinguishing CRC patients from controls with high accuracy (AUC = 0.86, 95% CI: 0.79–0.94). This signature was independently validated in our COLOBIOME cohort (AUC = 0.96, 95% CI: 0.92–1.00) and further tested in fecal immunochemical test (FIT) leftover samples, demonstrating compatibility with existing CRC screening workflows. Importantly, the signature effectively classified patients with early-stage tumors and advanced adenomas (AUC = 0.82, 95% CI: 0.71–0.97), underscoring its potential utility for early detection. Beyond its diagnostic relevance, our study provided additional novel insights into the biological role of miRNAs in CRC progression. Very importantly, we found that stool miRNA profiles mirrored those of tumor tissues, reinforcing thus their potential as biomarkers. Moreover, the detection of CRC-associated miRNA alterations in FIT leftover samples showed the feasibility of integrating stool miRNA analysis into routine screening programs.

Despite the study's strengths—such as its large sample size and rigorous multi-cohort validation—some challenges remain. CRC and adenoma subtypes were not exhaustively represented, and the investigation of miRNAs in screening samples remains in an early phase. Nevertheless, this work laid the foundation for refining non-invasive CRC diagnostics. By integrating stool miRNA profiling into existing screening strategies, we may significantly enhance early detection, improving patient outcomes while minimizing the need for invasive procedures.

Gene Expression Profiling of the Invasion Front for Risk Stratification in Stage IIA CRC

In a more applied extension of our work exploring CRC morphological heterogeneity, we hypothesized that focusing on the invasion front, rather than whole tumor sections, could provide additional insights for patient prognostic stratification. The tumor invasion front represents a biologically dynamic interface where cancer cells interact with the surrounding stroma, undergo epithelial-to-mesenchymal transition (EMT), and acquire invasive properties. Given its role in tumor progression, we investigated whether gene expression profiling of this specific region could improve risk assessment in Stage IIA CRC patients [***20***]. We specifically focused on Stage IIA CRC because this subgroup presents a major clinical challenge in treatment decision-making. While these patients generally have a good prognosis, a subset experiences early relapse despite the absence of traditional high-risk features. Unlike Stage III CRC, where adjuvant chemotherapy is standard, the benefit of additional treatment in Stage IIA remains debated.

We analyzed matched bulk tumor and invasion front samples from 39 patients, divided into early relapse (n = 19) and no relapse (n = 20) groups. While differential expression analyses did not reveal individual genes with significant differences after multiple testing corrections, pathway analyses highlighted the epithelial-to-mesenchymal transition (EMT) pathway as notably upregulated in early relapse cases. This finding underscores the invasion front's role in tumor aggressiveness.

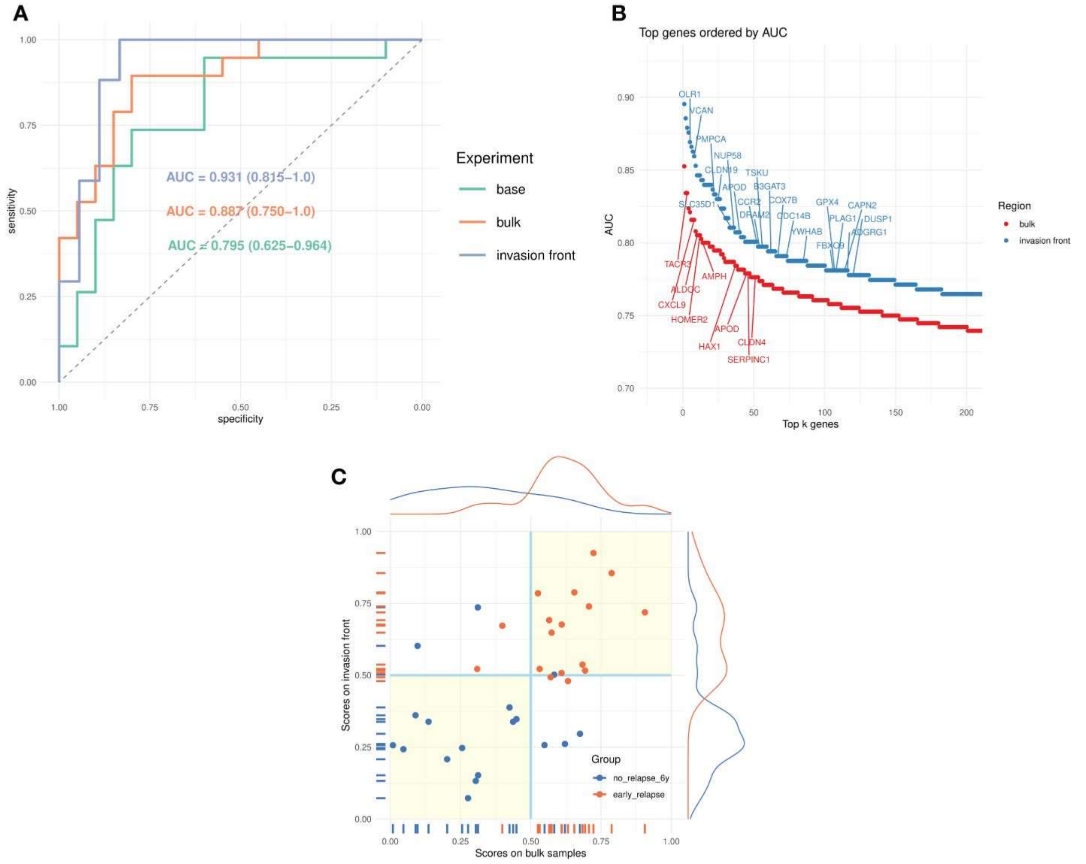


Figure 17. Prediction of early relapse. **(A)** Receiver operating characteristics (ROC) curves for the three models (baseline, bulk tumor, and invasion front) and the corresponding AUCs. **(B)** Univariate AUC, based on all samples, for top *k* = 200 genes from bulk tumor and invasion front expression profiles. The top genes (AUC > 0.775) from MSigDB hallmark signatures are marked. **(C)** Scatter plots of scores from bulk tumor and invasion front (35 samples) and their marginal distributions. The points are colored according to their true category, and the quadrants marked (light yellow background) indicate regions of agreement for the two classifiers (from [21])

By developing predictive models using ElasticNet regression, we discovered that gene expression profiles from the invasion front were more effective in forecasting early relapse than those from bulk tumor samples. The invasion front model achieved an area under the curve (AUC) of 0.931, surpassing the bulk tumor model's AUC of 0.887. This suggests that the invasion front harbors critical molecular insights pertinent to tumor progression.

4. CONCLUSION

Throughout my scientific journey, I have focused on integrating computational and molecular approaches to better understand the heterogeneity of colorectal cancer and its implications for clinical decision-making. My work has spanned multiple facets of CRC research, from mining large-scale molecular datasets to defining robust subtypes, linking these classifications to histopathological features, and validating key findings in preclinical models. The opportunity to work with the PETACC-3 clinical trial data allowed me to contribute to refining CRC molecular subtyping, particularly by identifying transcriptional programs associated with tumor location and progression. This naturally extended into investigating tumor morphology at a finer scale, where I explored how features at the invasion front contribute to patient risk stratification.

Recognizing that CRC is not just a tumor-intrinsic disease but one shaped by its microenvironment, I also turned my attention to the role of the microbiome. By leveraging our unique dataset of paired mucosal and stool samples, I contributed to characterizing microbial signatures associated with tumor subtypes and clinical variables. These findings not only deepen our understanding of CRC biology but also open possibilities for microbial-based biomarkers and therapeutic strategies.

A critical aspect of my work has been ensuring that computational insights are validated in biologically meaningful systems. I had the opportunity to contribute to efforts leveraging genetically engineered mouse models (GEMMs) and patient-derived xenografts (PDXs) to test hypotheses derived from our molecular analyses. This work underscored the challenges of translating in silico findings into preclinical models and the need for rigorous data integration strategies, some of which I helped shape.

Finally, my research has always been driven by its translational potential. From identifying biomarkers that predict response to chemotherapy to developing fecal microRNA signatures for noninvasive CRC diagnosis, I have sought to bridge the gap between discovery and clinical application. The ability to contribute to projects with real-world impact, including those within the ONCOBIOME consortium, has been particularly rewarding.

In summary, this thesis reflects my commitment to leveraging data-driven approaches to answer key questions in CRC research. By continuously refining methodologies and embracing interdisciplinary collaborations, I have aimed to contribute to a more precise and clinically actionable understanding of CRC. While many questions remain open, I see this work as a foundation for further exploration—both in the laboratory and in clinical practice.

5. REFERENCES

1. Dragomir MP, Popovici V, Schallenberg S, Čarnogurská M, Horst D, Nenutil R, et al. A quantitative tumor–wide analysis of morphological heterogeneity of colorectal adenocarcinoma [Internet]. 2024 [cited 2025 Feb 6]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.04.10.588907

2. International Agency for Research on Cancer. Global Cancer Observatory: Cancer Tomorrow (version 1.1). [Internet]. Available from: Available from: https://gco.iarc.who.int/tomorrow

3. Májek O, Zavoral M, Suchánek Š, Dušek L, Ngo O, Chloupková R, et al. Kolorektum.cz – Program kolorektálního screeningu v České republice [online] [Internet]. 2025. Available from: https://www.kolorektum.cz/cs/lekari/epidemiologie-a-vysledky-screeningu/

4. Linnekamp JF, Wang X, Medema JP, Vermeulen L. Colorectal Cancer Heterogeneity and Targeted Therapy: A Case for Molecular Disease Subtypes. Cancer Res. 2015 Jan 15;75(2):245–9.

5. Chan DKH, Buczacki SJA. Tumour heterogeneity and evolutionary dynamics in colorectal cancer. Oncogenesis. 2021 Jul 16;10(7):53.

6. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul;487(7407):330–7.

7. Schetter AJ, Okayama H, Harris CC. The role of microRNAs in colorectal cancer. Cancer J Sudbury Mass. 2012;18(3):244–52.

8. De Wit M, Fijneman RJA, Verheul HMW, Meijer GA, Jimenez CR. Proteomics in colorectal cancer translational research: Biomarker discovery for clinical applications. Clin Biochem. 2013 Apr;46(6):466–79.

9. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015 Nov;21(11):1350–6.

10. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. Stat Appl Genet Mol Biol. 2004 Jan 30;3(1):1–19.

11. R Core Team (2024). R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria.

12. Popovici V, Budinska E. Rgtsp [Internet]. Available from: https://github.com/bioinfo-recetox/Rgtsp

13. Schimek M, Budinska E, Kugler K. TopKLists on r-forge [Internet]. Available from: https://topklists.r-forge.r-project.org/

14. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004 Sep 15;5(10):R80.

15. Ivana Ihnatova EB. ToPASeq [Internet]. Bioconductor; 2017 [cited 2025 Feb 6]. Available from: https://bioconductor.org/packages/ToPASeq

16. Tejpar S, Yan P, Piessevaux H, Dietrich D, Brauchli P, Klingbiel D, et al. Clinical and pharmacogenetic determinants of 5-fluorouracyl/leucovorin/irinotecan toxicity: Results of the PETACC-3 trial. Eur J Cancer Oxf Engl 1990. 2018 Aug;99:66–77.

17. Brulé SY, Jonker DJ, Karapetis CS, O'Callaghan CJ, Moore MJ, Wong R, et al. Location of colon cancer (right-sided versus left-sided) as a prognostic factor and a predictor of benefit from cetuximab in NCIC CO.17. Eur J Cancer Oxf Engl 1990. 2015 Jul;51(11):1405–14.

18. Popovici V, Budinská E, Delorenzi M, Pavlicek A, Tejpar S, Weinrich SL. PROGNOSTIC AND PREDICTIVE GENE SIGNATURE FOR COLON CANCER [Internet]. US 61/470381, 61/413806; WO/2012/066451, PCT/IB2011/054962, 2012. Available from: https://patents.google.com/patent/WO2012066451A1/en

19. Vecchione L, Gambino V, Raaijmakers J, Schlicker A, Fumagalli A, Russo M, et al. A Vulnerability of a Subset of Colon Cancers with Potential Clinical Utility. Cell. 2016 Apr;165(2):317–30.

20. MoTriColor: A phase II study of vinorelbine in advanced BRAF-like colon cancer [Internet]. Netherlands Cancer Institute- Antoni van Leeuwenhoek hospital (NKI-AVL); 2018 Jan. Available from: https://www.clinicaltrialsregister.eu/ctr-search/search?query=2016-002364-13

21. Monti S. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach Learn. 2003;52(1/2):91–118.

22. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008 Mar 1;24(5):719–20.

23. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition: Molecular subtypes in colorectal cancer. Int J Cancer. 2014 Feb 1;134(3):552–62.

24. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med Genomics. 2012 Dec;5(1):66.

25. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LPMH, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med. 2013 May;19(5):614–8.

26. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization,

Validation, and Prognostic Value. Kemp C, editor. PLoS Med. 2013 May 21;10(5):e1001453.

27. Conti J, Thomas G. The role of tumour stroma in colorectal cancer invasion and metastasis. Cancers. 2011 Apr 26;3(2):2160–8.

28. Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat Genet. 2015 Apr;47(4):320–9.

29. Tuveson D, Hanahan D. Cancer lessons from mice to humans. Nature. 2011 Mar;471(7338):316–7.

30. Popovici V, Budinska E. WSItk: Whole Slide Imaging toolkit [Internet]. Available from: https://github.com/vladpopovici/WSItk

31. Colobiome: Tumour-adjacent microbiome and immune profile of tumour in the context of heterogeneity and aggressiveness of colorectal cancer [Internet]. Available from: https://starfos.tacr.cz/en/projekty/NV16-31966A?query=332iaacoohea

32. Morphogene [Internet]. Morphogene. Available from: https://morphogene.recetox.cz

33. Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. Expert Rev Anti Infect Ther. 2010 Apr;8(4):435–54.

34. Mantis NJ, Rol N, Corthésy B. Secretory IgA's complex roles in immunity and mucosal homeostasis in the gut. Mucosal Immunol. 2011 Nov;4(6):603–11.

35. Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, Medzhitov R. Recognition of Commensal Microflora by Toll-Like Receptors Is Required for Intestinal Homeostasis. Cell. 2004 Jul;118(2):229–41.

36. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol. 2012 Jun 25;10(8):575–82.

37. Vaupel P, Harrison L. Tumor hypoxia: causative factors, compensatory mechanisms, and cellular response. The Oncologist. 2004;9 Suppl 5:4–9.

38. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013 Aug 14;14(2):207–15.

39. Tlaskalová-Hogenová H, Štěpánková R, Kozáková H, Hudcovic T, Vannucci L, Tučková L, et al. The role of gut microbiota (commensal bacteria) and the mucosal barrier in the pathogenesis of inflammatory and autoimmune diseases and cancer: contribution of germ-free and gnotobiotic animal models of human diseases. Cell Mol Immunol. 2011 Mar;8(2):110–20.

40. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014 Oct;12(10):661–72.

41. Budinská E, Čarnogurská M. Mikrobiom v solidních nádorech. In: Tlaskalová-Hogenová, Helena a Eklová, Danka Mikrobiom a zdraví. Grada;

42. Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science. 2018 Jan 5;359(6371):91–7.

43. Derosa L, Hellmann MD, Spaziano M, Halpenny D, Fidelle M, Rizvi H, et al. Negative association of antibiotics on clinical activity of immune checkpoint inhibitors in patients with advanced renal cell and non-small-cell lung cancer. Ann Oncol. 2018 Jun;29(6):1437–44.

44. Geller LT, Barzily-Rokni M, Danino T, Jonas OH, Shental N, Nejman D, et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science. 2017 Sep 15;357(6356):1156–60.

45. Kasinskas RW, Forbes NS. Salmonella typhimurium lacking ribose chemoreceptors localize in tumor quiescence and induce apoptosis. Cancer Res. 2007 Apr 1;67(7):3201–9.

46. Sznol M, Lin SL, Bermudes D, Zheng L mou, King I. Use of preferentially replicating bacteria for the treatment of cancer. J Clin Invest. 2000 Apr 15;105(8):1027–30.

47. Westphal K, Leschner S, Jablonska J, Loessner H, Weiss S. Containment of Tumor-Colonizing Bacteria by Host Neutrophils. Cancer Res. 2008 Apr 15;68(8):2952–60.

48. Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Lett. 2014 Apr;345(1):17–20.

49. Tessem MB, Selnæs KM, Sjursen W, Tranø G, Giskeødegård GF, Bathen TF, et al. Discrimination of Patients with Microsatellite Instability Colon Cancer using^1 H HR MAS MR Spectroscopy and Chemometric Analysis. J Proteome Res. 2010 Jul 2;9(7):3664–70.

50. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. Cancer Prev Res (Phila Pa). 2014 Nov 1;7(11):1112–21.

51. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014 Nov;10(11):766.

52. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019 Apr;25(4):679–89.

53. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019 Apr;25(4):667–78.

54. Oncobiome.

6. LIST OF COMMENTED PUBLICATIONS

[***1***] Popovici V, **Budinska E,** Tejpar S, Weinrich S, Estrella H, Hodgson G, Roth AD, Bosman FT, Delorenzi M. Rgtsp: a generalized top scoring pairs package for class prediction. *Bioinformatics.* 2011 Mar 15;27(6):812-3. doi:10.1093/bioinformatics/btr003.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 12 2011, pages 1729–1730
doi:10.1093/bioinformatics/btr233

Data and text mining
Advance Access publication April 19, 2011

Rgtsp: a generalized top scoring pairs package for class prediction

Vlad Popovici[1,2,*], Eva Budinská[1,3] and Mauro Delorenzi[1]
[1]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, CH-1015 Lausanne, [2]Swiss National Center of Competence in Research Molecular Oncology, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland and [3]Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

Associate Editor: Jonathan Wren

ABSTRACT

Summary: A top scoring pair (TSP) classifier consists of a pair of variables whose relative ordering can be used for accurately predicting the class label of a sample. This classification rule has the advantage of being easily interpretable and more robust against technical variations in data, as those due to different microarray platforms. Here we describe a parallel implementation of this classifier which significantly reduces the training time, and a number of extensions, including a multi-class approach, which has the potential of improving the classification performance.

Availability and Implementation: Full `C++` source code and `R` package `Rgtsp` are freely available from http://lausanne.isb-sib.ch/~vpopovic/research/. The implementation relies on existing OpenMP libraries.

Contact: vlad.popovici@isb-sib.ch

Received on November 12, 2010; revised and accepted on April 8, 2011

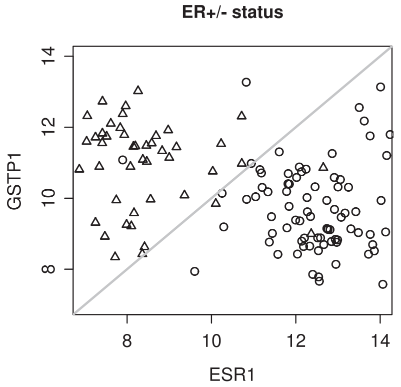


Fig. 1. Predicting estrogen receptor status: if GSTP1 < ESR1, then the sample is considered ER+ (circles), otherwise ER− (triangles).

1 INTRODUCTION

Top scoring pairs (TSPs; Geman *et al.*, 2004) are simple two-variables binary classifiers, in which the prediction of the class label is based solely on the relative ranking of the expression levels of the two genes. The rank-based approach to classification ensures a higher degree of robustness to technical variations and makes the rule easily portable across platforms. Also, the direct comparison of the expression level of the genes is easily interpretable in the clinical context, making the TSPs attractive for medical tests.

Let $\mathbf{x} = [x_i]_{i=1,\dots,m} \in \mathbb{R}^m$ be a vector of measurements (e.g. gene expression) representing a sample and let the corresponding class label be y, with two classes denoted by 0 and 1. Then, for all pairs of variables i and j, a score is computed,

$$s_{i,j} = P(x_i < x_j | y=1) - P(x_i < x_j | y=0), 1 \leq i,j \leq m \qquad (1)$$

where P are conditional probabilities estimated from the data, and the corresponding decision rule is: if $\text{sign}(s_{i,j})x_i < \text{sign}(s_{i,j})x_j$ then predict $y=1$, otherwise $y=0$. The pairs are ordered by the absolute values of their scores and the top t pairs ($t \geq 1$) are then considered for the final model (Geman *et al.*, 2004; Tan *et al.*, 2005; Xu *et al.*, 2005). Remarkably, training a TSP does not require the optimization of any parameter and does not depend on any threshold. Selecting a suitable value for t should be done following the usual machine learning paradigm for optimizing meta-parameters (see, for example, Hastie *et al.*, 2001). Figure 1 shows an example of a TSP predicting the estrogen receptor status. The decision boundary (in grey) is always a line with a slope of 1.

2 IMPLEMENTATION

While the method briefly described above is simple and poses no implementation problems, using it in the context of highly dimensional data requires the evaluation of an extremely large number of pairs of variables making its usage impractical, especially in the context of resampling techniques for performance estimation. However, most if not all of the modern desktop computers are multi-core machines, making parallel programs a feasible alternative to classical serial ones.

Our implementation in `C++` exploits the multi-core architecture by using the OpenMP libraries of the system (Chapman *et al.*, 2007), and is wrapped in an `R` package – `Rgtsp`. The full source code and the `R` package are available from http://lausanne.isb-sib.ch/~vpopovic/research/. As `C++` is the main implementation language, the library can easily be extended and integrated with other software libraries. Also, the `R` functions are independent of the domain of application so they could be applied to any kind of data.

*To whom correspondence should be addressed.

3 USAGE EXAMPLES

We present a typical case of using Rgtsp package. These examples represent solely some code snippets and not the full process of developing and assessing the performance of a classifier.

The data used in these examples consists of 130 samples stage I to III breast cancer (Hess *et al.*, 2006) and the goal is to predict the estrogen receptor status (positive or negative coded with '+1' and '0', respectively). For illustration purposes we use only a subset of full dataset available from GEO repository under accession number GSE16716.

Before starting R, the user has the option of choosing the number of processing units that will be used, by setting the environment variable OMP_NUM_THREADS. If not set, it defaults to the maximum number of processing units available.

The first steps load the library and the data and build a list of TSPs (note that the matrix *X* contains the variables as columns):

```
> library(Rgtsp)
> data(mdabr)
> tsp.list = tsp.n(X, y.erpos, 500)
> str(tsp.list)
> print(tsp.list)
```

The function tsp.n() returns at most *n* TSPs as a list with three components: the first two correspond to the indexes of the selected variables and the third one contains the associated scores. A similar function, tsp.s(), returns all the TSPs that have a score larger than a specified value.

For the *p*-th TSP, the prediction rule can be written as: predict class '+1' if X[,tsp.list$I[p]] < X[,tsp.list$J[p]] and this forms the core of the predict function. The decision function for $p=1$ in the above example is shown in Figure 1. Given a list of TSPs one has different choices on how to obtain the final predicted labels. Currently, Rgtsp proposes two means of combining the predictions of individual TSPs: either by majority voting or by weighting the votes with the correspoding scores—giving more weight to the TSPs with better scores. This functionality is available through the predict() generic function:

```
> yp = predict(tsp.list, X, combiner="majority")
> sum(yp != y.erpos) # count the errors
[1] 3
```

By inspecting the list of TSPs, it becomes clear that there are variables that are selected many times as having always either higher or lower value than all its pairing variables. We call such a structure a *TSP hub* and we can construct all the hubs larger than a specified size (25 pairs for example) using

```
> h = tsp.hub(tsp.list, min.hub.size=25)
> print(h)
Hub 1: 194 pairs
    Center: 953 >
14 25 42 43 44 45 54 105 140 146 149 150 152 202 ...
```

This corresponds to a TSP hub in which the probeset colnames(X)[953] (205225_at, ESR1) has a higher expression than all other probesets in the list tsp.list. The TSP hubs can also be used in predicting the labels, through the same mechanism as above:

```
> yph = predict(h, X, combiner="majority")
> sum(yph != y.erpos) # no. of errors: 6
```

We see that in this particular case the prediction by TSP hubs is slightly less accurate than the combined predictions of the individual TSPs.

The generalization performance of the TSPs classifiers can be estimated by various methods. The Rgtsp package provides a function for *k*-fold cross-validation of the binary TSP classifiers (either tsp.n() or tsp.s() functions), cv.tsp(), which returns the training and validation performance of the classifier (it defaults to 5-fold cross-validation).

```
> r = cv.tsp(X, y.erpos)
> print(r)
$tr.m
 Error.rate Sensitivity Specificity         AUC
 0.02884615  0.97812500  0.96000000  0.96906250
```

In the case of a multi-class problem, we propose to use classification trees built on top of TSPs predictions. For $C>2$ classes, one can train TSPs to solve each of the $C(C-1)/2$ pairwise binary classification problems [called one-versus-one (Hsu and Lin, 2002) or round robin (Fürnkranz , 2002) strategy] and then combine the predictions of the TSPs through a classification tree to predict the original classes. For more details the reader is referred to the package web page. This approach is implemented in the function mtsp() and makes use of the ctree() function in the party R package (y4 is an artificial 4–class label vector):

```
> m = mtsp(X, y4)
> yp = predict(m, X)
```

Funding: Swiss National Science Foundation NCCR Molecular Oncology (to V.P. and M.D.); Fondation Medic (to E.B.).

Conflict of Interest: none declared.

REFERENCES

Chapman,B. *et al*. (2007) *Using OpenMP*. The MIT Press.

Fürnkranz,J. (2002) Round robin classification. *J. Mach. Learn. Res.*, **2**, 721–747.

Geman,D. *et al*. (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article19.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer.

Hess,K.R. *et al*. (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, **24**, 4236–4244.

Hsu,C.-W. and Lin,C.-J. (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.*, **13**, 415–425.

Tan,A.C. *et al*. (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896–3904.

Xu,L. *et al*. (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **21**, 3905–3911.

[*2*] Schimek MG, **Budinská E**, Kugler KG, Švendová V, Ding J, Lin S. TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. Stat Appl Genet Mol Biol. 2015 Jun;14(3):311-6. doi: 10.1515/sagmb-2014-0093. PMID: 25968440.

Software and Application Note Open Access

Michael G. Schimek*, Eva Budinská, Karl G. Kugler, Vendula Švendová, Jie Ding and Shili Lin

TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists

Abstract: High-throughput sequencing techniques are increasingly affordable and produce massive amounts of data. Together with other high-throughput technologies, such as microarrays, there are an enormous amount of resources in databases. The collection of these valuable data has been routine for more than a decade. Despite different technologies, many experiments share the same goal. For instance, the aims of RNA-seq studies often coincide with those of differential gene expression experiments based on microarrays. As such, it would be logical to utilize all available data. However, there is a lack of biostatistical tools for the integration of results obtained from different technologies. Although diverse technological platforms produce different raw data, one commonality for experiments with the same goal is that all the outcomes can be transformed into a platform-independent data format – rankings – for the same set of items. Here we present the R package TopKLists, which allows for statistical inference on the lengths of informative (top-*k*) partial lists, for stochastic aggregation of full or partial lists, and for graphical exploration of the input and consolidated output. A graphical user interface has also been implemented for providing access to the underlying algorithms. To illustrate the applicability and usefulness of the package, we integrated microRNA data of non-small cell lung cancer across different measurement techniques and draw conclusions. The package can be obtained from CRAN under a LGPL-3 license.

Keywords: 62G99; 65K10; 68N01; 65C60; 62F07.

DOI 10.1515/sagmb-2014-0093

1 Introduction

Several high-throughput technologies have emerged in the past decade, most notably next generation sequencing, but also methods that estimate abundance levels of proteins and small molecules. Together, these methods are contributing to an enormous collection of experimental data. However, current research in molecular science is typically based on rather small studies in terms of sample size, many of them addressing the same disease or target. The findings obtained across platforms and studies are often quite diverse and an

***Corresponding author: Michael G. Schimek,** Statistical Bioinformatics, IMI, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria, e-mail: michael.schimek@medunigraz.at
Eva Budinská: Bioinformatics in Translational Research, IBA, Masaryk University, Kotlarska 2, 61137 Brno, Czech Republic
Karl G. Kugler: Institute for Bioinformatics and Systems Biology, Helmholtz Centre Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
Vendula Švendová: Statistical Bioinformatics, IMI, Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria
Jie Ding: Stanford Cancer Institute, Stanford University, 265 Campus Drive, Stanford, CA 94305-5456, USA
Shili Lin: Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

increasingly important task is to strengthen the evidence of these findings. Hence, there is a strong demand for statistical methods that integrate such findings, for example for combining microarray-based expression measurements with RNA-seq results.

A central task is the integration of such data, which differ in important aspects such as laboratory technology, quantification, scale, and study size. When several studies are combined, the involved sets of genes or of other omics entities usually do not match and missing observations are likely to occur. Moreover, often only subsets of unknown size of these data are relevant or informative. In almost all situations the original metric measurements from the involved studies can be transformed into rank data. Until recently, most integration tools for rank data have been heuristic in nature and could not meet all the above mentioned demands. The few statistical integration approaches in use are limited to microarray results (Yang et al., 2006; Plaisier et al., 2010). A general methodology allowing for the integration of other high-throughput technologies, as well as allowing for a platform and technology mix, even when ranked lists are incomplete, had been lacking until the work of Lin and Ding (2009) and Hall and Schimek (2012). Schimek et al. (2012) combined these approaches and extended them with the goal of processing arbitrarily long multiple ranked lists. To turn such novel statistical methods into practical tools, we have implemented them in the TopKLists R package. It focuses on the nonparametric estimation of the top-*k* list length and on the stochastic aggregation of the identified top-*k* lists. In addition, it also includes conventional aggregation techniques and visual aids for the analysis of ranked lists and the interpretation of aggregation results. In the following, we give an overview of the package and its statistical background, and we apply it to microRNA lung cancer data obtained from a number of different platforms.

2 Structure and availability of the R package

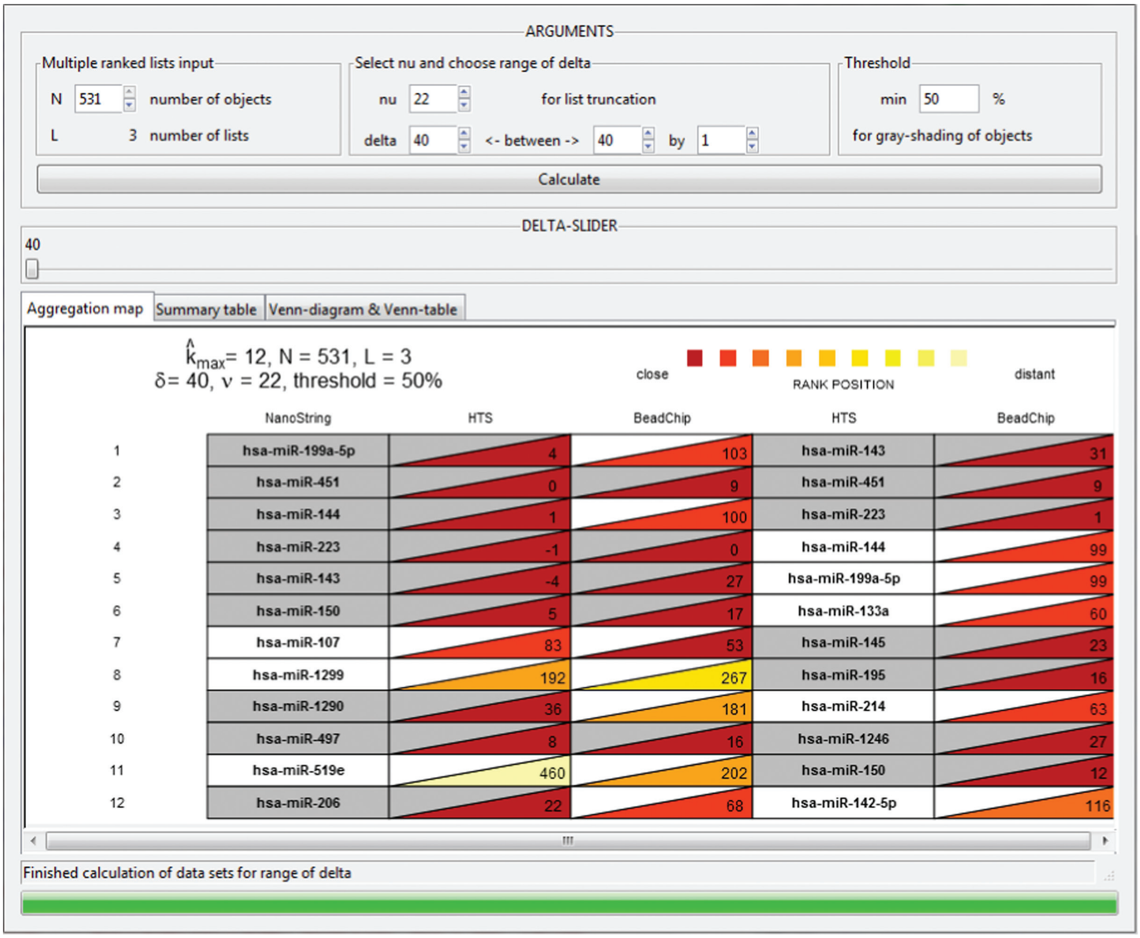
The TopKLists package comprises three modules: (i) TopKInference offers exploratory nonparametric inference for the estimation of the top-*k* list length of paired rankings; (ii) TopKSpace provides various rank aggregation techniques; (iii) TopKGraphics comprises a collection of graphical tools for the exploration of data and for the visualization of aggregation results. The analysis pipeline is to estimate the top-*k* consensus list length first, which also works for more than two ranked lists comprising tens of thousands of items, and to then aggregate the already obtained truncated lists. A new graphical concept, the aggregation map, has been implemented to visualize this graphically. It displays the selected top items with quality measures indicating their relevance with respect to the full ranked lists. Venn-type representations and a summary list form the end of the pipeline. The obtained formal results can then be used in succeeding downstream analysis and experimental validation. The modules can be used as stand-alone R libraries or via a graphical user interface (GUI) for ease of use (see Figure 1 for an example of the GUI interface).

TopKLists is available under the LGPL-3 license from CRAN for all major operating systems. Its R-Forge Web page http://TopKLists.r-forge.r-project.org/ offers the latest development version of the package, detailed vignette-based information about the methods and the package, and instructions on how to analyze the application data described in the example of this paper.

3 Implementation and performance of the R package

The TopKLists package has been designed and implemented for usage on standard desktop computers. To increase the computational speed and performance, parts of the sampling methods have been implemented in C. Therefore, when locally building the package from the source code these methods will be compiled. The graphical user interface, which provides interactive access to several TopKLists' procedures, has been implemented using the gWidgets2 package (Verzani, 2014).

The time needed for computation in the modules TopKInference and TopKSpace depends strongly on the choice of tuning parameters (see next section). Typically, when these parameters are chosen



ARGUMENTS

Multiple ranked lists input
N 531 number of objects
L 3 number of lists

Select nu and choose range of delta
nu 22 for list truncation
delta 40 <- between -> 40 by 1

Threshold
min 50 %
for gray-shading of objects

DELTA-SLIDER
40

$\hat{k}_{max}$= 12, N = 531, L = 3
δ= 40, ν = 22, threshold = 50%

close 🟥 🟥 🟧 🟧 🟧 🟨 🟨 🟨 🟨 distant
RANK POSITION

	NanoString	HTS	BeadChip	HTS	BeadChip
1	hsa-miR-199a-5p	4	103	hsa-miR-143	31
2	hsa-miR-451	0	9	hsa-miR-451	9
3	hsa-miR-144	1	100	hsa-miR-223	1
4	hsa-miR-223	-1	0	hsa-miR-144	99
5	hsa-miR-143	-4	27	hsa-miR-199a-5p	99
6	hsa-miR-150	5	17	hsa-miR-133a	60
7	hsa-miR-107	83	53	hsa-miR-145	23
8	hsa-miR-1299	192	267	hsa-miR-195	16
9	hsa-miR-1290	36	181	hsa-miR-214	63
10	hsa-miR-497	8	16	hsa-miR-1246	27
11	hsa-miR-519e	460	202	hsa-miR-150	12
12	hsa-miR-206	22	68	hsa-miR-142-5p	116

Finished calculation of data sets for range of delta

Figure 1: GUI window with the final aggregation map of the NSCLC application.

appropriately, for lists with thousands of items, the runtime should be in the range of several seconds for inference as well as for aggregation. Most of the implemented aggregation techniques are computationally more demanding than the inference approach. For this reason, aggregation is usually performed on partial lists obtained from the inference procedure. For stochastic aggregation techniques runtime can amount to five or more seconds in a typical scenario.

4 Brief description of the statistical methods

The purpose of the `TopKInference` module is inference on the concordant top length of several rankings comprising the same set of items. The assumptions are: the reliability of rankings breaks down after the first *k* items due to lack of discriminatory information, irregular or even missing assessments, and substantially more ranked items than assessors exist. The index j_0 is the rank position where the consensus information of two lists degenerates into noise. The estimation of $\hat{j}_0-1=\hat{k}$ is achieved via a moderate deviation-based method developed by Hall and Schimek (2012).

For a given set of items, the input is the overlap of rank positions represented by a sequence of indicators, where I_j=1 if the ranking, given by the second assessor to the item ranked *j* by the first assessor, is not more than δ index positions distant from *j*, otherwise I_j=0. The assumption that the variables I_j follow a Bernoulli random distribution can be relaxed. There is theoretical and simulation evidence that dependencies among the ranked lists do not impair the estimates (Hall and Schimek, 2012). As well as the *distance* δ, the inter-assessor or inter-platform variability, there is another tuning parameter, the *pilot sample size* ν, which is a smoothing parameter controlling the irregularity of assessments or expression measurements. A graphical method called Δ-plot is implemented in TopKGraphics which helps to select δ. The parameter ν can be chosen interactively via the GUI.

The overall estimate $\hat{k}_{max}$ for ℓ multiple lists is calculated in the following way: The inference method is applied to all pairwise list combinations $\mathscr{L}=(\ell^2-\ell)$ of the lists, thus we obtain $\mathscr{L}$ values $\hat{k}_j$ (j=1, 2, ..., $\mathscr{L}$). The overall top-k list length is then defined by $\hat{k}_{max}=\max_j(\hat{k}_j)$ (note, other criteria could be chosen as well). The ℓ full lists truncated after $\hat{k}_{max}$ form the input to TopKSpace. As the reader will see from the description below, TopKSpace is more general, and this specific scenario constitutes a special case that TopKSpace is applicable to.

The principle of the TopKSpace module is to consolidate information from the ℓ top-k lists to arrive at an aggregate list, AL. The top-k lists (L_1, L_2, ..., L_ℓ) may not only be of different lengths, they may also come from studies or assessments that consider different sets of items, hence the underlying spaces (S_1, S_2, ..., S_ℓ) from which the top-k lists are derived may actually be different. The goal therefore is to find the top-k list, AL, from the aggregate new space $(\cup_{i=1}^{\ell} L_i)$, such that the weighted sum of distances between each of the input lists and AL will be the minimum among lists of the same length. Two distance measures, Kendall's τ and Spearman's footrule, are available in the package. Both take the differences in the underlying spaces into account (Lin, 2010). There are three classes of algorithms implemented in TopKSpace, namely Borda's method, Markov chain (MC) algorithms (Lin, 2010), and a cross entropy Monte Carlo (CEMC) method taking advantage of the new order explicit algorithm (OEA) as described by Lin and Ding (2009). The Borda and MC methods consist of heuristic algorithms that do not directly optimize the objective function (i.e., minimizing the weighted distances), whereas the CEMC method employs a Monte Carlo search procedure for achieving this optimization. Borda and MC algorithms run substantially faster than the CEMC algorithm, however the latter usually achieves better results. Nevertheless, simulation studies indicate that taking the underlying space into consideration has a much greater impact than using different algorithms.

5 Application to cross platform microRNA profiles

Stimulated by the methodological discussion of microRNA profiling in Baker (2010), we compared non-small cell lung cancer (NSCLC) cell lines grown in vitro and in vivo as xenograft models across platforms. From the NCBI GEO database we retrieved data (Tam et al., 2014) of five in vitro and five in vivo samples from three different platforms: (i) GSE51501, Illumina Human v2 MicroRNA Expression BeadChip; (ii) GSE51504, NanoString nCounter Human v1 miRNA Expression Assay; (iii) GSE51507, Illumina HiSeq 2500 (High Throughput Sequencing, abb. HTS). Data (i) and (ii) were normalized using Bioconductor's normalize.quantiles and analysed with the R-package samr (Tibshirani et al., 2011) (cell line vs. xenograft). The next generation sequencing data (iii) were processed with Bioconductor's DESeq2 (Love et al., 2014). The resulting miRNA expression values (items) from each platform were ranked according to their FDR-adjusted p-values. Those items common to all three lists were the input to the package TopKLists and comprise N=531 miRNAs. The thus obtained ranked lists and the corresponding code for the data analysis can be accessed and downloaded from the TopKLists Web page.

Data exploration led to the choice of δ=40 and ν=22 for the inference procedure (for details please refer to the show case instructions on the Web page). The obtained result was $\hat{k}_{max}=12$ and the three lists were truncated at this index position. The associated aggregation map is displayed in Figure 1. Its left group (*NanoString-HTS-BeadChip*) represents the aggregation result when all three platforms are integrated, and the right group (*HTS-BeadChip*) when *NanoString* is excluded. A group comparison allows us to identify platform differences ('white' denotes that an item is top-listed in only one of the concerned lists, 'gray' otherwise). NanoString had the strongest impact on the selection of top-ranking miRNAs and forms, with the other two platforms, a highly conforming group of six items. hsa-miR-107, on rank 7 in NanoString, is of special interest, as it was shown to suppress growth of NSCLC cell lines and induced a G1 cell cycle arrest in H1299 cells (Takahashi et al., 2009). It was ranked 83 and 53 index positions away in HTS and BeadChip, and therefore is represented as *close* in the map by a 'red' color as opposed to more *distant* ranks, presenting themselves in 'yellow'.

Finally, we calculated an optimized aggregate list $\widehat{AL}$ for the three lists truncated after $\hat{k}_{max}=12$ via CEMC under Kendall's τ and Spearman's footrule. In Table 1 (columns 1 and 2) the final items are displayed in their new rank order. For comparison, the 12 top-ranked miRNAs based on Fisher's method for combining *p*-values (Fisher, 1925) are listed in the third column of the same table. We have used the function `fisher.method` from the `R` package `MADAM` (Kugler et al., 2010) with Benjamini-Hochberg *p*-value correction.

The CEMC stochastic search algorithm may select items that are top-ranked only in one of the lists (here *BeadChip*). This applies to the following items in Table 1: hsa-miR-576-5p, hsa-miR-490-5p, hsa-miR-139-5p, hsa-miR-1233, hsa-miR-1284, and hsa-miR-505. In contrast, Fisher's method tends to select 'consensus' items, thus having greater agreements with the aggregation map results. Within the top-5 positions the same items are selected by all methods. Only the orders are permuted. However, apart from this rather limited set of overlapping miRNAs, both aggregate lists from CEMC, as well as the aggregation map discussed before, clearly point at substantial platform differences.

Using the miRSystem (Lu et al., 2012) we found the final lists (one for Kendall, one for Spearman, and one for Fisher's method) of ranked miRNAs to be highly enriched for the JAK-STAT signaling pathway and the Hedgehog signaling pathway both of which were suggested to play an important role in NSCLC. The interesting candidates comprise hsa-miR-143, which is among a set of 43 miRNAs that were found to be differentially expressed between noncancerous lung tissues and lung cancer tissues (Yanaihara et al., 2006) and has also been suggested as a putative biomarker for NSCLC (Gao et al., 2010). Finally, on rank 1 and on rank 2, respectively, we have the RAB14 targeting tumor suppressor hsa-miR-451 (Wang et al., 2011).

6 Discussion

A major advantage over ground truth-based and other ad hoc methods is `TopKLists`'s ability to provide an objective data-driven top-list length estimate and a consolidated as well as optimized aggregate ranking based on multiple input lists. In the described NSCLC application it allowed us to efficiently select those miRNAs which are supported by all three or at least by two platforms. In addition, a consolidated set of miRNAs under different aggregation criteria (distance measures) could be obtained. The aggregation map

Table 1: Aggregate list results of the NSCLC application.

Rank	$\widehat{AL}$ (Kendall)	$\widehat{AL}$ (Spearman)	Fisher's method
1	**hsa-miR-451**	**hsa-miR-143**	**hsa-miR-143**
2	**hsa-miR-223**	**hsa-miR-451**	**hsa-miR-451**
3	**hsa-miR-199a-5p**	**hsa-miR-223**	**hsa-miR-223**
4	**hsa-miR-143**	**hsa-miR-144**	**hsa-miR-144**
5	**hsa-miR-144**	**hsa-miR-199a-5p**	**hsa-miR-199a-5p**
6	**hsa-miR-150**	hsa-miR-1284	**hsa-miR-145**
7	hsa-miR-576-5p	hsa-miR-139-5p	**hsa-miR-133a**
8	hsa-miR-490-5p	**hsa-miR-150**	**hsa-miR-195**
9	hsa-miR-139-5p	**hsa-miR-195**	**hsa-miR-214**
10	**hsa-miR-107**	**hsa-miR-145**	**hsa-miR-150**
11	hsa-miR-1233	hsa-miR-505	**hsa-miR-1246**
12	**hsa-miR-133a**	**hsa-miR-1246**	**hsa-miR-142-5p**

First and second columns: CEMC consolidated list results under the distance measures Kendall's τ and Spearman's footrule. Third column: consolidated list using Fisher's method for combining *p*-values (miR-symbols in **bold** coincide with the aggregation map result in Figure 1).

as well as the stochastic CEMC aggregation method also aid in giving an answer to the problem raised in Baker (2010): Although there is high conformity among the top-5 items across all (graphical and stochastic) aggregation techniques, our results support the observation that substantial platform differences exist with respect to all other miRNA measurements. As has been demonstrated in this paper, `TopKLists` offers a variety of highly useful and computationally efficient state-of-the-art methods for omics data integration, most of them implemented in `R` for the first time.

Acknowledgments: Gratefully acknowledged are financial support of the Medical University of Graz (MGS, VS) as well as funding by the Deutsche Forschungsgemeinschaft (DFG) to SFB 924 (KGK) and by the US National Science Foundation grant DMS-1220772 (SL).

References

Baker, M. (2010): “MicroRNA profiling: separating signal from noise,” Nat. Methods, 7, 687–692.

Fisher, R. A. (1925): Statistical methods for research workers, Edinburgh: Oliver and Boyd, ISBN 0-05-002170-2.

Gao, W., Y. Yu, H. Cao, H. Shen, X. Li, S. Pan and Y. Shu (2010): “Deregulated expression of miR-21, miR-143 and miR-181a in non small cell lung cancer is related to clinicopathologic characteristics or patient prognosis,” Biomed Pharmacother, 64, 399–408.

Hall, P. and M. G. Schimek (2012): “Moderate-deviation-based inference for random degeneration in paired rank lists,” J. Am. Stat. Assoc., 107, 661–672.

Kugler, K. G., L. A. Mueller and A. Graber (2010): “MADAM: an open source meta-analysis toolbox for R and Bioconductor,” Source Code Biol. Med., 5, 3.

Lin, S. (2010): “Space oriented rank-based data integration,” Stat. Appl. Genet. Mol. Biol., 9:Article20. doi: 10.2202/1544-6115.1534. Epub 2010 Apr 9.

Lin, S. and J. Ding (2009): “Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies,” Biometrics, 65, 9–18.

Love, M. I., W. Huber, and S. Anders (2014): “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” Genome. Biol., 15:550.

Lu, T.-P., C.-Y. Lee, M.-H. Tsai, Y.-C. Chiu, C. K. Hsiao, L.-C. Lai and E. Y. Chuang (2012): “miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets,” PloS One, 7, e42390.

Plaisier, S. B., R. Taschereau, J. A. Wong and T. G. Graeber (2010): “Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures,” Nucleic Acids Res., 38, e169.

Schimek, M. G., A. Myšicková and E. Budinská (2012): “An inference and integration approach for the consolidation of ranked lists,” Commun. Stat. – Simul. C., 41, 1152–1166.

Takahashi, Y., A. R. Forrest, E. Maeno, T. Hashimoto, C. O. Daub and J. Yasuda (2009): “MiR-107 and MiR-185 can induce cell cycle arrest in human non small cell lung cancer cell lines,” PloS One, 4, e6677.

Tam, S., R. de Borja, M.-S. Tsao, and J. D. McPherson (2014): “Robust global microRNA expression profiling using next-generation sequencing technologies,” Lab. Invest., 94, 350–358.

Tibshirani, R., G. Chu, B. Narasimhan and J. Li (2011): “Significance analysis of microarrays – samr: R package version 2.0,” URL http://CRAN.R-project.org/package=samr.

Verzani, J. (2014): “gWidgets: gWidgets API for building toolkit-independent, interactive GUIs – R package version 0.0-54,” URL http://CRAN.R-project.org/package=gWidgets.

Wang, R., Z. Wang, J. Yang, X. Pan, W. De and L. Chen (2011): “MicroRNA-451 functions as a tumor suppressor in human non-small cell lung cancer by targeting ras-related protein 14 (RAB14),” Oncogene, 30, 2644–2658.

Yanaihara, N., N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, et al. (2006): “Unique microRNA molecular profiles in lung cancer diagnosis and prognosis,” Cancer Cell, 9, 189–198.

Yang, X., S. Bentink, S. Scheid and R. Spang (2006): “Similarities of ordered gene lists,” J Bioinform Comput Biol., 4, 693–708.

[3] Ihnatova I, **Budinska E**. ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. *BMC Bioinformatics*. 2015; 16:350. doi:10.1186/s12859-015-0763-1.

BMC Bioinformatics

SOFTWARE Open Access

ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data

Ivana Ihnatova[1*] and Eva Budinska[1,2,3]

Abstract

Background: Pathway analysis methods, in which differentially expressed genes are mapped to databases of reference pathways and relative enrichment is assessed, help investigators to propose biologically relevant hypotheses. The last generation of pathway analysis methods takes into account the topological structure of a pathway, which helps to increase both specificity and sensitivity of the findings. Simultaneously, the RNA-Seq technology is gaining popularity and becomes widely used for gene expression profiling. Unfortunately, majority of topological pathway analysis methods remains without implementation and if an implementation exists, it is limited in various factors.

Results: We developed a new R/Bioconductor package ToPASeq offering uniform interface to seven distinct topology-based pathway analysis methods, of which three we implemented de-novo and four were adjusted from existing implementations. Apart this, ToPASeq offers a set of tailored visualization functions and functions for importing and manipulating pathways and their topologies, facilitating the application of the methods on different species. The package can be used to compare the differential expression of pathways between two conditions on both gene expression microarray and RNA-Seq data. The package is written in R and is available from Bioconductor 3.2 using AGPL-3 license.

Conclusion: ToPASeq is a novel package that offers seven distinct methods for topology-based pathway analysis, which are easily applicable on microarray as well as RNA-Seq data, both in human and other species. At the same time, it provides specific tools for visualization of the results.

Keywords: Topology, Pathway analysis, Microarray, RNA-Seq, Packages

Background

High-throughput gene expression technologies (such as microarray or RNA-Seq) are used to estimate expression levels of thousands of genes in one experiment. Often the aim of such experiments is to find pathways and biological processes altered between two conditions, which helps investigators to propose biologically relevant hypotheses for further research. Achieving this aim implies integration of a priori known pathway information into the data analysis. Most often, a set of genes with similar biological function or participating in a regulatory process is employed as a set of entities in enrichment-based methods [1]. This approach, however, ignores known interactions between particular genes reflected in the topological structure. Thus, if a change in interactions occurs, this is not reflected in the results. The last generation of pathway analysis methods takes into account the topological structure of a pathway, which helps to increase both specificity and sensitivity of the findings.

Several types of methods for topology-based pathway analysis were proposed in the recent years (for review see [2]) - in all of them, the topological structure of a pathway is represented as graph with nodes (genes, proteins) and edges (interactions between genes/proteins). The methods test one of the two types of null hypotheses as proposed in [3] for gene set enrichment analysis.

*Correspondence: ihnatova@iba.muni.cz
[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Brno, Czech Republic
Full list of author information is available at the end of the article

Independently on the hypothesis tested, we can further distinguish *multivariable* and *univariable* methods. For more detailed description of differences between multivariable vs univariable methods, we refer the reader to Additonal file 1.

Here, we focus on methods that (i) aim to identify pathways affected between two conditions based on differential expression of genes in the pathway - the most frequent aim of high-throughput genomic data studies, (ii) use the a priori known pathway topologies and (iii) use the pathway topologies separately.

The vast majority of existing topology pathway analysis methods were designed for continuous gene expression measures as obtained from microarray experiments. In order to apply them to discrete count data - a typical output from RNA-Seq experiment (number of reads mapped to a particular gene) - one must use a suitable transformation. Poisson or Negative binomial distribution are used as model distributions in differential expression analysis at gene-level for RNA-Seq data and a wide range of both transformation methods and statistical tests for this purpose exists. Performance of these methods is only recently being compared in extensive simulation studies [4–7].

The published methods are only rarely implemented as a publicly available software tool or package, and sometimes the existing implementation is not available anymore (e.g TAPPA [8]). The existing implementations can be divided into three categories: (i) commercial products (e.g. MetaCore [9]); (ii) R-packages (e.g. SPIA [10]) (iii) standalone applications (e.g. PWEA [11] or PRS [12]) and (iv) web-based applications (e.g. iPathwayGuide [13]). All of these tools usually offer embedded pathway topologies with a limited battery of methods (typically only one) and simple visualization (if any) of the results. Simultaneous application of different methods and comparison of their results is therefore very time-consuming, cumbersome and prone to clerical errors due to need for repeated data conversion and transfer. Additionally, the results may not be directly comparable, since some of the implementations use either built-in pathway topologies or their own pathway topology processing algorithm that leads to different topological structures. One of the best existing tools offering common interface to four topology-based pathway analysis methods (TopologyGSA [14], clipper [15], DEGraph [16] and SPIA [17]) is the R/Bioconductor package `graphite` [18]. The user can also access lists of parsed pathway topologies for some of the most common experimental organisms (14 in version 1.14.1) from several distinct databases (up to 6 for H. Sapiens, same version) stored as objects of class `PathwayList` where individual pathways are represented as instances of class `Pathway`. Although more pathways can be obtained from public databases or specialized websites and parsed to the R environment with available CRAN/Bioconductor packages, there is no transformation function from other pathway classes to the `PathwayList` or `Pathway`. The current `graphite` implementation has no uniform way of calling methods or specification of their parameters, making simultaneous application of different methods unhandy. Additionally, SPIA is limited only to data with EntrezGene identifiers and the signs of the interactions are neglected in DEGraph.

Here, we present ToPASeq (Topology-based Pathway Analysis of microarray and RNA-Seq data) - a Bioconductor package that adjusts the set of methods available through `graphite` and extends them by addition of three more methods. The package offers their unified manipulation and provides tools for their easy application on RNA-Seq count data. In addition, it provides special functions for conversion of user-imported pathways into `Pathway` class and a set of tools for coercing graphs between different formats and manipulation and visualization of the results.

In section Implementation, we describe the software implementation and available functions. Concrete examples of package usage and its comparison to other tools are given section Results and Discussion.

Implementation

ToPASeq was implemented using statistical programming language R and the package is available through the open-source Bioconductor project [19].

In order to apply a topology-based pathway analysis method we need (i) gene expression measurements (a gene expression data matrix in which rows refer to genes and columns to samples), (ii) a vector with sample class labels and (iii) a list of pathways of interest together with their topologies in a specific format. The gene expression measurements and sample class information are usually available from the experiment.

Pathway topologies and their manipulation

Pathway topologies are necessary for topology-based pathway analysis and can be created manually, or - even better - obtained from public databases or R packages, where they are typically stored in one of the standardized formats (KGML, BioPax, specific R classes). These formats, however, need to be parsed (downloaded and converted to specific format) to be used within the methods' particular implementations. Within R framework, multiple ways exist for pathway topology/graph representation. More detailed description of some of them in the context of biological pathways can be found in Additional file 1.

Our package requires the pathway topologies in format defined as S4 class `PathwayList` where individual pathways are of class `Pathway`, which allows combination of oriented and not-oriented edges as well as multiple edges between nodes. We have especially designed several transformation functions that convert the most common formats into `Pathway`.

The users might be interested in manual editing of topology of the parsed pathways. We added group of methods such as (i) adding/removing of the nodes and edges, (ii) changing the type of interaction/directionality, (iii) merging two pathways into one, (iv) obtaining the induced subgraph. Additionally, the user may need to select only a subset of pathways based on their topological properties (e.g. number of edges related to a particular node, number of nodes, number of edges, number of connected components etc.). These can be easily obtained with other set of available functions.

Moreover, we especially designed a new function `reduceGraph` which merges the user defined named sets of nodes into a single node. The members of the sets must form either a gene family or a protein complex. The another function `estimateCF` estimates the maximal list of the sets of the nodes that can be merged. Finally, we provide a general function `convertIdentifiersByVector` which requires user specified information. For the detailed desctiption of the functionalities mentioned above we refer the reader to Additional file 1.

Methods for topology-based pathway analysis

The package offers seven different methods covering various approaches in topological pathway analysis (see Table 1 for details). For detailed description of each method the reader is referred to cited references. We will focus on those aspects that are relevant to methods' new implementation. All methods are implemented as a single function that applies the method over the list of pathways. More detailed description of differences between previous implementations of methods to our implementation can be found in Additional file 1.

We imported and adjusted the implemetation of the following methods: TopologyGSA, DEGraph, SPIA and Clipper. We found that the original implementation of the TopologyGSA method is extremely computationally intense for some of the pathways as the authors employ function that implements the exact branch-and-bound algorithm [20] to detect all of the cliques (subsets of nodes where every two nodes are connected by an edge) in a pathway topology. In our implementation, we substituted this function with `getCliques` which implements more efficient Bron-Kerbosch algorithm [21]. For the DEGraph method we have created a new wrapper function that preserves the possibility to consider interaction types (activation and inhibiton) and transforms the results into more user-friendly format - a data frame. The previous implementations of the SPIA method were limited to Entrez identificators. In our package we have bypassed this limitation by incorporating a more general converting function. Additionally, the user can also obtain a gene-level net perturbation accumulation — a measure of the importance of a gene in the topology. The Clipper method constists of two steps: (i) first, the differential expression of a pathway is assessed, (ii) then, the pathway topology is transformed into a junction tree and the portions of the tree which are mostly associated with phenotype are identified. We designed a new function that performs both steps of the algorithm in a single call.

In all of the imported and adjusted implementations we also added, when appropriate, an additional parameter specifying how should be the undirected interactions oriented. The user can choose whether an edge is oriented in both directions or only in one according to the order of the nodes.

We de-novo implemented three methods: TAPPA, PWEA, PRS, for which there was no implementation available within R framework. The PRS and PWEA are implemented in MATLAB and C++ respectively and these tools are discussed in the section Comparison with other Tools. Our de-novo implementations are settled for pathway topologies from `graphite` package where one node is represented by only one gene or protein. Both PWEA and PRS methods incorporate a permutation-based test in order to assess the statistical significance of the pathway

Table 1 Methods included in the package

Method	Ref.	Type[a]	Hypothesis	A/I[b]	Primary Graph	Implementation	Input data[c]
TopologyGSA	[14]	M	self-contained	No	DAG	adjusted	GEDM
DEGraph	[16]	M	self-contained	Yes	DAG	adjusted	GEDM
clipper	[15]	M	self-contained	No	DAG	adjusted	GEDM
SPIA	[17], [25]	U	competitive	Yes	directed	adjusted	DEG and their log fold-change
PRS	[26]	U	competitive	No	directed	de novo	DEG and their log-fold change
PWEA	[27]	U	competitve	No	undirected	de novo	gene-level statistics
TAPPA	[8]	U	self-contained	No	undirected	de novo	GEDM

[a] - M - multivariable, U - univariable [b] - A - Activation, I - Inhibition [c] - the data related to the pathway topology

score. Considering the computational complexity of this approach we parallelized the crucial step of the PWEA method (repeated application of the differential expression analysis). In addition, the function for obtaining the number of the differentially expressed genes in PRS algorithm was implemented in C++ via `Rcpp` package.

While several methods (TopologyGSA, DEGraph, Clipper and TAPPA) work directly with normalized gene expression values, others (SPIA, PRS and PWEA) use the result of differential gene-expression analysis with or without application of significance thresholds to obtain the list of differentially expressed genes (Fig. 1). With respect to this, all the methods were adapted also for a simple use of RNA-Seq count data. First, we employed pre-processing step for RNA-Seq normalization, with a selection of two best performing methods TMM [22], DESeq [23], as compared in Dillies et al. [4] and regularized log transformation from `DESeq2` package which effectively removes the mean-variance relationship known in RNA-Seq data. Second, we added methods for RNA-Seq differential gene expression analysis (from `limma` and `DESeq2` packages).

Usage and visualization

Each method is implemented as a single wrapper function which allows the user to call a method in a single command. The wrapper function offers: (i) normalization of count data; (ii) differential gene expression analysis and (iii) pathway analysis. The data input types were unified for all the methods. Expression data can be supplied both as matrix or as `ExpressionSet`. The functions' outputs have uniform format defined as a new S3 class `topResult` with specified output of generic functions (print, plot, summary) and methods for accessing individual slots of the resulting object. The users can specify which method should be used for normalization or differential expression analysis of the RNA-Seq data, with respect to their own preferences. This data pre-processing step can be completely omitted and users can submit already normalized data or, if appropriate, the results of the differential expression analysis (a table containing log fold-changes, statistics and *p*-values). Note, that PWEA method requires also so called Topology Influence Factors (TIFs), which need to be calculated from normalized gene expression data matrix.

When the generic function `plot()` is applied to a `topResult` class, together with a name of the pathway or position in the list of pathways identifying the pathway to be plotted, a visualization of the pathway with three gene-level statistics is produced (Fig. 1 in Additional file 1. The user can specify a threshold by which an agreement between the expression status of the nodes and the interaction type between them is examined (Fig. 2 in Additional file 1).

The topology can be reduced by user specified list of nodes that are to be merged into one node. In this situation a pie chart is used as a representation of a node and the number of slices equals to the number of nodes merged. The filling colour and the radius is preserved from the separated nodes (Fig. 2). By default a mean change of the gene expression is used as a representative of the values when the agreement between gene expression and the interaction type is examined, but the user can specify another aggregation function. A slightly modified graph is plotted for TopologyGSA and Clipper, which perform differential expression analysis of the cliques. Since

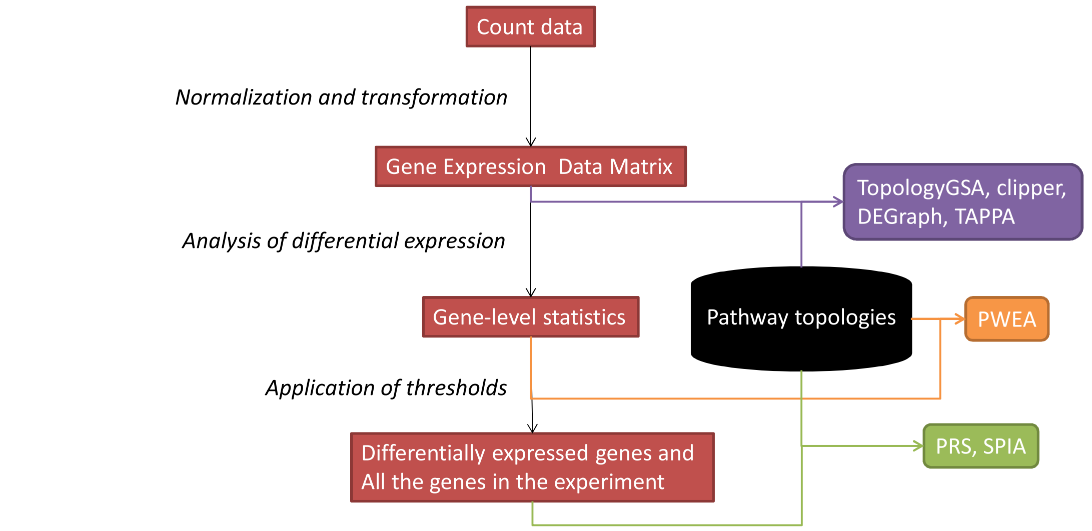

Count data
Normalization and transformation
Gene Expression Data Matrix
TopologyGSA, clipper, DEGraph, TAPPA
Analysis of differential expression
Gene-level statistics
Pathway topologies
PWEA
Application of thresholds
PRS, SPIA
Differentially expressed genes and All the genes in the experiment

Fig. 1 Schema of a processing pipeline. The red boxes refer to the outputs from regular analysis of differentially expressed genes and possible inputs for topology-based pathway analysis. Arrows indicate the processing pipeline of each of the methods implemented in the package

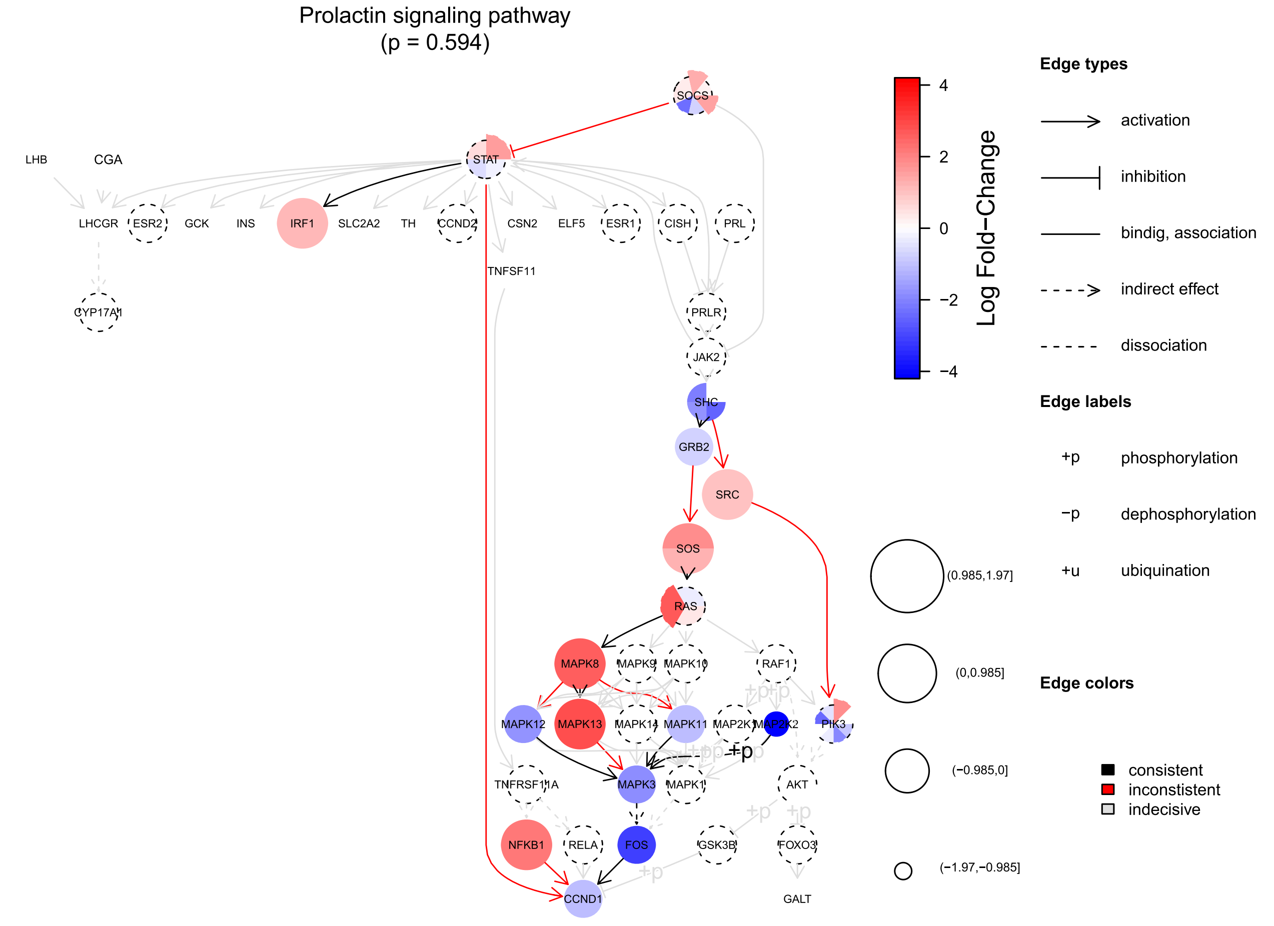


Fig. 2 Visualization of the results after merging some of the gene families into one node. Some of the genes families present in the pathway were merged into single nodes. Those nodes are drawn as pie-chart, in which the number of slices equals to the number of gene merged. The colour, border and radius are preserved from the complete graph (Fig. 2 in Additional file 1). Average log fold-change is used as representative value, when the agreement between expression and interaction type is assessed

a single node can be a member of more than one clique, the colour of edges is used for their visualization (Fig. 4 in Additional file 1).

Results and Discussion

For a simple example of how to create and manipulate a pathway, we refer the reader to Additional file 1.

We provide a simple application example of implemented methods on a RNA-Seq dataset. For more detailed descriptions of all the functions we refer the reader to the package manual.

The aim is to compare gene expression profiles between wild-type and RNA-binding protein hnRNP C (HNRNPC) knockdown HeLa cells [24]. The RNA-Seq dataset came from `gageData` package. There are four knockdown samples and four experimental samples in this dataset containing the count data for 22932 genes. We load the data and remove genes with count 0 in all samples:

```
> library(ToPASeq)
> library(gageData)
> data(hnrnp.cnts)
> group<-c(rep("sample",4),
  rep("control",4))
> hnrnp.cnts<-hnrnp.cnts[rowSums
  (hnrnp.cnts)>0,]
```

... download the KEGG pathways and apply all seven topology-based pathway methods:

```
> kegg<-pathways("hsapiens","kegg")
> top<-TopologyGSA(hnrnp.cnts, group, kegg,
  type="RNASeq")
> deg<-DEGraph(hnrnp.cnts, group, kegg,
  type="RNASeq")
> cli<-clipper(hnrnp.cnts, group, kegg,
  type="RNASeq")
> spi<-SPIA(hnrnp.cnts, group, kegg,
  type="RNASeq")
```

Table 2 Known implementation of the methods provided in ToPASeq

Method	Language	Source	Pathways	Format	Input data	Methods	Issusses
`topologyGSA`	R	Bioconductor	one example	`graphNEL`	GEDM	topologyGSA	too computationaly intense
`clipper`	R	Bioconductor	imported from `graphite`	`pathway`	GEDM	clipper	two separate steps necessary
`DEGraph`	R	Bioconductor	parsing function for KGML	`graphNEL`	GEDM	DEGraph	
`SPIA`	R	Bioconductor	parsing function for KGML, H. sapiens and M. musculus pre-parsed	list of adjacency matrices	DEG and log fold-changes	SPIA	Only for EntrezGene IDs
PRS tool	MATLAB	web[a]	KEGG	unknown	GEDM	PRS	can not add or modify pathways, the data must have manufacturer probeset IDs, limited set of: possible platforms, DE tests,
PWEA	C++	web[b]	human pathways from KEGG	unknown	GSD	PWEA	only for UNIX-like
TAPPA	Java	web[c]	KEGG or PPI added to a gene set	-	-	TAPPA	not available
`graphite`	R	Bioconductor	pathways for 14 species from up to 6 databases	`Pathway`	depends on the method	topologyGSA, clipper, SPIA, DEGraph,	suboptimal import of the methods

[a] - http://www.buckingham.ac.uk/research/clore-laboratory-diabetes-obesity-and-metabolic-research/staff/maysson-al-haj-ibrahim/prs-tool/

[b] - http://zlab.bu.edu/PWEA/index.php

[c] - http://watson.mcgee.mcw.edu:8080/~sgao, the page is down. (First accessed 4 Apr 2012) PPI - protein-protein interactions GEDM - gene expression data matrix, log2-transformed and normalized expression profiles

```
> prs<-PRS(hnrnp.cnts, group, kegg,
  type="RNASeq")
> pwea<-PWEA(hnrnp.cnts, group, kegg,
  type="RNASeq")
> tap<-TAPPA(hnrnp.cnts, group, kegg,
  type="RNASeq")
```

The arguments of all functions are as follows (from left to the right): a count matrix (or gene expression data matrix), a grouping vector, list of pathways with topologies and a type of the data). The TMM normalization and the `limma`-based differential gene-expression analysis are used by default. The pre-set thresholds for considering a gene significant are p-value less than 0.05 and the absolute log fold change above 2. Further, the gene identifiers in pathways are automatically converted to the EntrezGene identifiers and the non-oriented edges are oriented in both directions, when required.

The results for an individual pathway can be visualized as shown in Fig. 1 in Additional file 1:

```
> plot(spi,"Prolactin signaling pathway",
+ kegg, fontsize=50)
```

Comparison with other tools

The known previous implementations of the methods (if any) offered in ToPASeq are summarized in Table 2. We will further discuss only the methods implemented de-novo in R/Bioconductor frame work. For TAPPA there is no other available implementation known to the authors. A C++ implementation of PWEA can be downloaded from http://zlab.bu.edu/PWEA/download.php. The expression data have to be in the GSD format from Gene Expression Omnibus, where the probesets are named by both manufacturer IDs and the gene symbols. It is coupled with python script for retrieving and processing of KEGG .xml and .gene files. Beside the limitation to KEGG pathways and the need for manual downloading of non-human pathways or conversion to KGML format, it can be run only on UNIX-like systems. Recently, a standalone MATLAB-based implementation of PRS was published [12]. The application requires normalized microarray data in XLS file with manufacturer identifiers of the probesets, together with specification of the platform and the normalization method that was applied to the data. The set of possible platforms is limited to selection of Affymetrix HG and one Agilent platform. The user has no control over the pathway topologies that are used.

None of these tools allows for different method for normalization (e.g normalization with custom CDF-files from http://brainarray.mbni.med.umich.edu) or differential expression analysis; nor can it be used to analyse the RNA-Seq data.

Some users may prefer Cytoscape for visualization of pathways, since it provides user-friendly and interactive interface, which can be achieved using the `RCytoscape` package. Within this interface, however, the user can specify only the basic graphical parameters like size, shape or colour of the nodes or the styles of edges. Advanced graphical approaches provided through plug-ins can be accessed only directly from Cytoscape. We are currently working on the option of interactive graph visualization.

Conclusions

Topology-based pathway analysis comprises a new generation of methods in gene set analysis, with the potential of generating more sensitive and more specific results. Currently, high-throughput technologies producing gene expression data that serve as input to these methods are employed in almost every biological and biomedical research with RNA-Seq being in the leader position. Tools for comfortable and quick application of these methods and visualization of their results are needed. Available packages or standalone applications are usually limited to one or few methods, readily applicable mainly to human studies and rarely contain also a visualization tool. We propose `ToPASeq`, a Bioconductor package providing a set of easy-to-use and general tools for topology-based pathway analysis within the R workspace. It offers seven distinct topology-based pathway analysis methods that cover wide range of approaches and can be easily applied on both microarray and RNA-Seq data. It also offers a visualization tool that is able to capture all the relevant information about the expression of genes within one pathway. Finally, the functions for pathway conversion extend the application of topology-based pathway analysis to experiments on species other than human.

Availability and requirements

Project name: ToPASeq
Project home page: http://www.bioconductor.org/packages/release/bioc/html/ToPASeq.html
Operating system(s): Platform independent
Programming language: R
Other requirements: R version 3.2.1, CRAN and Bioconductor packages: graphite (>= 1.14), graph, gRbase
License: AGPL-3
Any restrictions to use by non-academics: none
Availability of supporting data: EBI ArrayExpress Experiment E-MTAB-1147: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1147/, also in `gageData` package

Additional file

Additional file 1: Supplementary material.pdf. The file contains additional details on the following: i) common principles of the multivariable and univariable topology-based methods; ii) the functions for pathway creation and manipulation (desciption as well as demostration); iii) comparison of ToPASeq with existing tools. (1013 Kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

II and EB designed the package. II implemented the methods in R and drafted the manuscript. All authors contributed to, read and approved the final version of the manuscript.

Acknowledgements

Part of the work was carried out with the support of Proteomics Core Facility of CEITEC - Central European Institute of Technology, ID number CZ.1.05/1.1.00/02.0068, financed from European Regional Development Fund. This research was supported by RECETOX research infrastructure and by the Czech Ministry of Education (LM2011028, LO1214).

Author details

[1]Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Brno, Czech Republic. [2]Central European Institute of Technology, Brno, Czech Republic. [3]RECETOX, Faculty of Science, Masarykova Univerzita, Brno, Czech Republic.

Received: 7 September 2015 Accepted: 7 October 2015
Published online: 29 October 2015

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–15550.
2. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol. 2013;4(278):1–22.
3. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinforma. 2007;23(8):980–7.
4. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
5. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of rna-seq data. BMC Bioinforma. 2013;14(1):91.
6. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome Biol. 2013;14(9):95.
7. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in rna-seq studies. Brief Bioinforma. 2015;16(1):59–70.
8. Gao S, Wang X. Tappa: topological analysis of pathway phenotype association. Bioinforma. 2007;23(22):3100–102.
9. Thomson R. MetaCoreTM Data-mining and Pathway Analysis. http://thomsonreuters.com/metacore/. Access Date: 13 Jul 2013.
10. Tarca AL, Kathri P, Draghici S. SPIA: Signaling Pathway Impact Analysis (SPIA) Using Combined Evidence of Pathway Over-representation and Unusual Signaling Perturbations, R package version 2.16.0. 2013. http://bioinformatics.oxfordjournals.org/cgi/reprint/btn577v1. Access Date: 10 Sep 2013.
11. Hung JH. PWEA Pathway Enrichment Analysis. http://zlab.bu.edu/PWEA/index.php. Access Date: 13 Jul 2014.
12. Ibrahim M, Jassim S, Cawthorne MA, Langlands K. A matlab tool for pathway enrichment using a topology-based pathway regulation score. BMC Bioinforma. 2014;15:358.
13. Advaita C. iPathwayGuide. http://www.advaitabio.com/products.html. Access Date: 13 Jul 2013.
14. Massa M, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. BMC Syst Biol. 2010;4(1):121.
15. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 2013;41(1):e19.
16. Jacob L, Neuvial P, Dudoit S. Gains in Power from Structured Two-Sample Tests of Means on Graphs: Annals of Applied Statistics; 2012. 6:pp. 561–600.
17. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-s, et al. A novel signaling pathway impact analysis. Bioinforma. 2009;25(1):75–82.
18. Sales G, Calura E, Cavalieri D, Romualdi C. graphite - a bioconductor package to convert pathway topology to gene network. BMC Bioinforma. 2012;13(1):20.
19. Gentleman RC, Carey VJ, Bates DM. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:80.
20. Niskanen S, Östergård PRJ. Cliquer user's guide, version 1.0. Technical report. Espoo, Finland: Communications Laboratory, Helsinki University of Technology; 2003.
21. Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph. Commun ACM. 1973;16(9):575–7.
22. Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010;11(3):25.
23. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106.
24. Luo W, Friedman M, Shedden K, Hankenson K, Woolf P. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinforma. 2009;10(1):161.
25. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17(10):000.
26. Al-Haj Ibrahim M, Jassim S, Cawthorne MA, Langlands K. A topology-based score for pathway enrichment. J Comput Biol. 2012;19(5): 563–573.
27. Hung JH, Whitfield T, Yang TH, Hu Z, Weng Z, DeLisi C. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol. 2010;11(2):23.

[***4***] Ihnatova I, Popovici V, **Budinska E**. A critical comparison of topology-based pathway analysis methods. *PLOS ONE.* 2018;13(1):e0191154. doi:10.1371/journal.pone.0191154.

RESEARCH ARTICLE

A critical comparison of topology-based pathway analysis methods

Ivana Ihnatova[1,2], **Vlad Popovici**[1], **Eva Budinska**[1,2]*

1 RECETOX, Faculty of Science, Masarykova Univerzita, Brno, Czech Republic, **2** Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Brno, Czech Republic

* budinska@recetox.muni.cz

OPEN ACCESS

Citation: Ihnatova I, Popovici V, Budinska E (2018) A critical comparison of topology-based pathway analysis methods. PLoS ONE 13(1): e0191154. https://doi.org/10.1371/journal.pone.0191154

Editor: Xia Li, College of Bioinformatics Science and Technology, CHINA

Received: March 12, 2017

Accepted: December 29, 2017

Published: January 25, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by the CETOCOEN PLUS project (CZ.02.1.01/0.0/0.0/15 003/0000469), by the Ministry of Health of the Czech Republic under grant no.16-31966A and by the European Community's Seventh Framework Programme under grant agreement no. 602901 MErCuRIC. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research,

Abstract

One of the aims of high-throughput gene/protein profiling experiments is the identification of biological processes altered between two or more conditions. Pathway analysis is an umbrella term for a multitude of computational approaches used for this purpose. While in the beginning pathway analysis relied on enrichment-based approaches, a newer generation of methods is now available, exploiting pathway topologies in addition to gene/protein expression levels. However, little effort has been invested in their critical assessment with respect to their performance in different experimental setups. Here, we assessed the performance of seven representative methods identifying differentially expressed pathways between two groups of interest based on gene expression data with prior knowledge of pathway topologies: SPIA, PRS, CePa, TAPPA, TopologyGSA, Clipper and DEGraph. We performed a number of controlled experiments that investigated their sensitivity to sample and pathway size, threshold-based filtering of differentially expressed genes, ability to detect target pathways, ability to exploit the topological information and the sensitivity to different preprocessing strategies. We also verified type I error rates and described the influence of overexpression of single genes, gene sets and topological motifs of various sizes on the detection of a pathway as differentially expressed. The results of our experiments demonstrate a wide variability of the tested methods. We provide a set of recommendations for an informed selection of the proper method for a given data analysis task.

Introduction

High-throughput gene expression technologies (microarrays or next-generation sequencing) allow the estimation of the expression levels of thousands of genes in a single experiment. Often these experiments are just a first step in a broader biological investigation and serve generating hypotheses based on identified differentially expressed genes and pathways. A biological pathway is a collection of genes or molecules that act synergistically by means of chemical reactions, molecule modifications or signal transduction to execute a biological function. Thus, from a computational analysis perspective, a pathway is a set of genes (proteins) and their associated pairwise interactions. Pathway analysis aims to discover those pathways whose

Development, and Innovations Infrastructures". The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

activation/inactivation is associated with a group of interest. This type of analysis requires integrating information about gene ontology and pathway structure.

Generally, there are two main approaches: one that relies only on the expression levels of the constituent genes (of the pathway)—and is epitomised by the GSEA family of methods—and a second one that additionally exploits the pathway topology. The second group of methods represents a more recent evolution of pathway analysis methods that try to improve both specificity and sensitivity of the findings.

The application of topology-based methods is facilitated by the existence of public databases which gather information about gene/protein interactions, such as the well-known Kyoto Encyclopedia of Genes and Genomes (KEGG) database which provides access to hundreds of pathways representing state-of-the-art knowledge about molecular interactions. Prior to performing a topology-based pathway analysis, the pathway of interest must be pre-processed into a simple interaction network.

Each new topology-based pathway method usually compares its performance to an enrichment-based method (most often GSEA [1]) on a set of benchmark datasets. Sometimes, the underlying mathematical model is verified by simulations. The reviews that include topology-based pathway analysis methods either examine their algorithms from mathematical perspective [2–4] or their performance on both real and simulated data [5, 6]. The latter approach revealed that topology-based methods outperform enrichment-based methods in accuracy and sensitivity only for non-overlapping pathways [5] and that the FCS variant of CePa [7] method exhibits the best cross-study concordance [6]. However, there are multiple limitations to the existing comparisons which hamper the identification of actionable information about the most appropriate method for a given analytical problem. First, the comparison of a topology-based method with enrichment-based methods is oversimplistic as it does not investigate the topological aspects of pathway deregulation (position and biological importance of a gene in a pathway, deregulation of topological motifs etc.). Second, the existing reviews do not examine the effect of pathway topology pre-processing strategy or whether the inclusion of the pathway topology information in the analysis has actually any effect at all. Third, multiple other effects, such as sample size (crucial aspect in biological experiments) or the effect of a deregulation of a single or very few genes, are not explored either.

Given the proliferation of methods (see [8] for a review of 22 methods) and with limited insight into their performance, data analysts are confronted with the difficult task of selecting the best-suited method for analysing the data at hand. We propose a systematic investigation of several prominent recently proposed methods and provide a simple guideline for decision-making.

In this work, we consider a number of parameters that influence the quality of the results obtained by topology-based pathway analysis. These parameters are varied in controlled experiments in order to study the sensitivity of the methods and—when possible—to quantify it. These experiments are performed on both artificial and real-world data, thus resulting in a comprehensive characterisation of the behaviour of each considered method. From the beginning, we did not expect to identify a single method that would fit all possible applications, thus, in our investigations, we tried to capture most of the standard scenarios. The methods under investigation were selected based on the following criteria: (i) the aim is to detect differentially expressed pathways (DEPs) between two groups of interest based on gene expression data; (ii) the pathway topology is a priori known and is modeled as simple interaction network or graph $G = (V, E)$, where V is a set of vertices/nodes represented by products of genes and E is a set of edges representing interactions between them; (iii) the pathways are modeled and analyzed individually (without cross-pathway interactions). The typical input data for these methods consists of a gene expression data matrix (log2-transformed normalised expression profiles from a high-throughput technology after standard pre-processing), group membership labels (as a vector) and the list of pathway topologies. Based on these criteria we selected the following methods: SPIA [9], PRS [10], CePa [7], TAPPA [11], TopologyGSA [12], Clipper [13] and DEGraph [14]. Each method assigns a test-statistic and a *p*-value to each pathway (possibly other parameters like the number of differentially expressed genes, pathway size etc.) and pathways with extreme test-statistic or low *p*-value are called'differentially expressed'.

Materials and methods

We performed eight distinct experiments to provide comprehensive insight into the topology-based pathway analysis methods (Fig 1, Table 1, S1 Text). In these experiments, we examined the influence of the number of parameters on the results obtained by topology-based pathway analysis methods. A detailed description of the experiments can be found in the S1 Text.

The first group of parameters are data set-centric (sample size, pathway size, number of DEGs in the dataset and thresholds used to detect DEGs; Experiment 1) and helped us to describe the performance of a method under various conditions and to guide the selection of

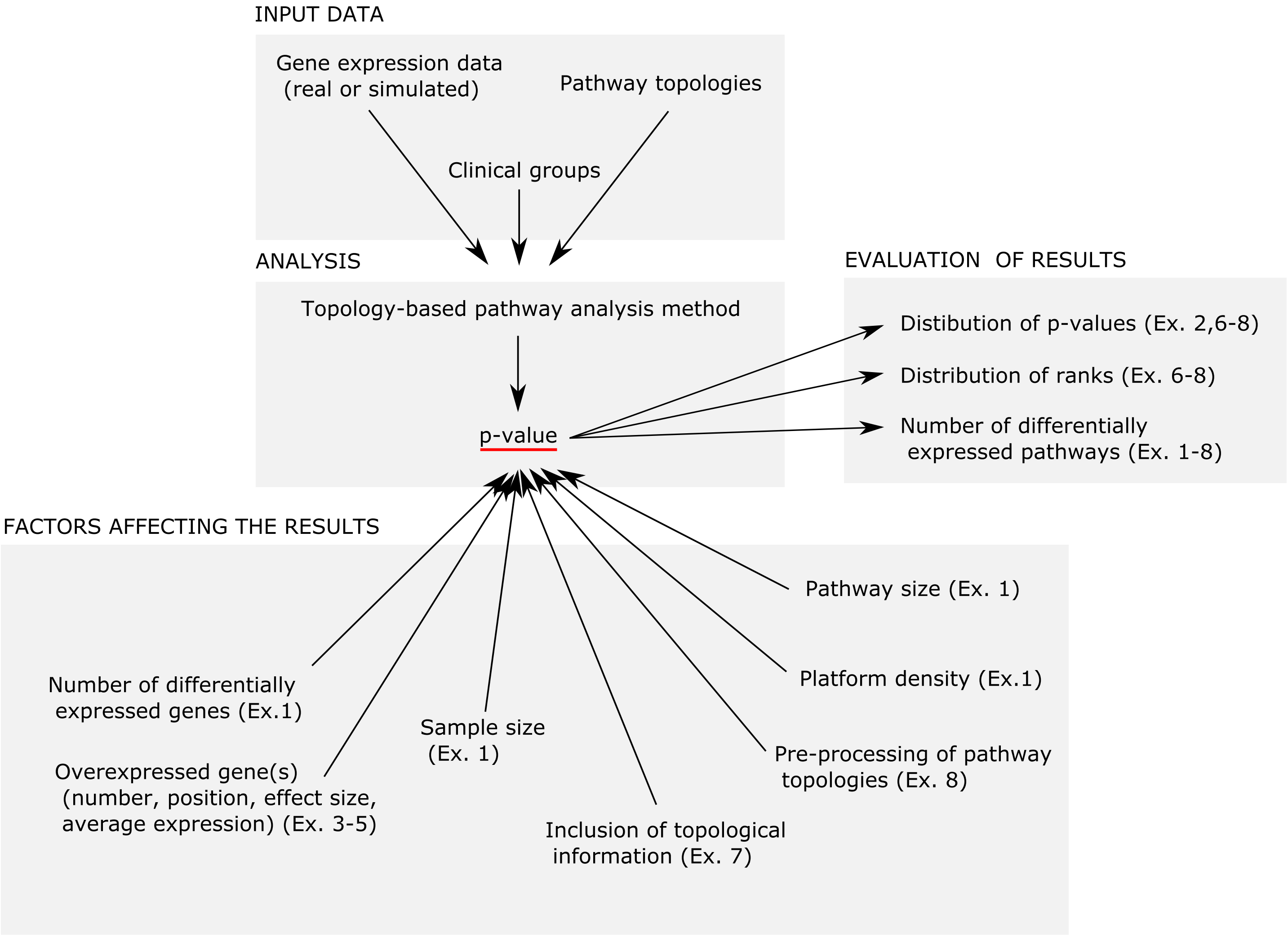

INPUT DATA
Gene expression data (real or simulated)
Pathway topologies
Clinical groups
ANALYSIS
Topology-based pathway analysis method
p-value
EVALUATION OF RESULTS
Distibution of p-values (Ex. 2,6-8)
Distribution of ranks (Ex. 6-8)
Number of differentially expressed pathways (Ex. 1-8)
FACTORS AFFECTING THE RESULTS
Pathway size (Ex. 1)
Number of differentially expressed genes (Ex.1)
Platform density (Ex.1)
Sample size (Ex. 1)
Pre-processing of pathway topologies (Ex. 8)
Overexpressed gene(s) (number, position, effect size, average expression) (Ex. 3-5)
Inclusion of topological information (Ex. 7)

Fig 1. Overview of the eight controlled experiments (Ex. 1-8) performed.

https://doi.org/10.1371/journal.pone.0191154.g001

Table 1. Overview of the experiments performed to evaluate methods' performance.

Experiment	Parameter(s) under study	Varied parameter(s)*	Datasets	Pathway topologies	Evaluation criterion †
1	Effect of sample size, pathway size and significance thresholds for DEGs	n_1, n_2, $\|V\|$, θ	Simulated, Real	`graphite`	Prop. DEPs
2	Type I error rate	y	Simulated	`graphite`	Prop. DEPs, histogram
3	Single gene overexpression	X_{ij}, $i \in I \subset V$, $\|I\| = 1$, $j \in 1, 2, 3, \ldots, n$ such that $y_j = 1$	Simulated	`graphite`	Prop. DEPs
4	Multiple genes overexpression	X_{ij}, $i \in I \subset V$, $\|I\| \in 2, 3, 4, 5$, $j \in 1, 2, 3, \ldots, n$ such that $y_j = 1$	Simulated	`graphite`	Prop. DEPs
5	Topological motif overexpression	X_{ij}, $i \in I \subset V$, $\|I\| \in 3, 4, 5$, $j \in 1, 2, 3, \ldots, n$ such that $y_j = 1$	Simulated	`graphite`	Prop. DEPs
6	Target pathway detection	X_{ij}, $i \in I \subset V$, $\|I\| = 1$, $j \in 1, 2, 3, \ldots, n$ such that $y_j = 1$	Simulated, Real	`graphite`	Median p-value, rank
7	Inclusion of topological information	PT	Simulated, Real	`graphite` ‡	Prop. DEPs
8	Pre-processing of pathway topologies	PT	Simulated, Real	`ToPASeq`	Prop. DEPs

*X is a normalized $\log_2$-transformed gene expression data matrix of expression profiles of p genes (rows) and $n_1 + n_2$ samples (columns), n_1 and n_2 denote number of samples in two compared groups, y is a vector of 1's and 2's assigning samples into the groups, PT is a set of pathway topologies (graphs) $G = (V, E)$, where V is a set of vertices/nodes represented by products of genes and E is a set of edges representing interactions between them, θ is the threshold used for detection of DEGs;

†Prop. DEPs denotes Proportion of Differentially Expressed Pathways;

‡without interactions

https://doi.org/10.1371/journal.pone.0191154.t001

the optimal method for a specific dataset. The methods' ability to control type I error was studied in Experiment 2. The influence of overexpression of particular gene(s) (Experiments 3-5), the influence of discarding the topological information (Experiment 7) and the effect of the pre-processing of pathway topologies (Experiment 8) tested the topology-based nature of the methods. If no effects were observed, the method should not be considered as topology-based pathway analysis method. The increased sensitivity and specificity expected from the incorporation of the topological information were examined by the identification of biologically relevant pathways (Experiments 6-8) since no proper method for identifying truly differentially expressed pathways is known.

Following the categorization of GSEA methods, the topology-based pathway analysis methods can be grouped based on three main criteria: (i) the null hypothesis (*competitive* and *self-contained*); (ii) the (non)identification of differentially expressed genes (DEGs) prior pathway analysis (*over-representation analysis (ORA)* and *functional class scoring (FCS)*) and (iii) the number of variables in the model (*univariable* and *multivariable*) (see S1 Text for the details). We will use these categories in methods evaluation.

For each experiment we applied selected methods (Table 2) on gene expression datasets, looking for differentially expressed pathway(s) between two groups of interest from a collection of pathways. In ORA methods we detected differentially expressed genes with moderated t-test [15] and significance level $\alpha = 0.05$, unless stated otherwise. For all methods estimating significance threshold using permutations, the number of permutations was set to 1000. The pathways were considered differentially expressed if their p-value was below the significance threshold $\alpha = 0.05$. All the analyses were performed in R statistical framework [16] and Bioconductor [17]. There are multiple freely-available implementations of the selected topology-based pathway analysis methods: (i) original implementation (all but TAPPA), (ii) `graphite`

Table 2. Overview of the selected methods.

	SPIA	PRS	CePa	TAPPA	TopologyGSA	Clipper	DEGraph
Reference	[9, 20, 21]	[10]	[7]	[11]	[12]	[13]	[14]
Null hypothesis	C	C	C	*	SC	SC	SC
ORA/FCS	ORA	ORA	ORA	FCS	FCS	FCS	FCS
Type	U	U	U	U	M	M	M
Pathway model	DG	DG	UG, DG	UG	DAG	DAG	UG
Node statistic	Log FC	Log FC	Log FC	-	-	-	-
Topology usage	Perturbation factor	Downstream DEG	Centrality	PCI	GGM, IPS	GGM, IPS	GL, FT
Pathway statistic	Impact factor	Sum	Sum	*	T^2	T^2	T^2
Statistical significance	Gene perm.	Gene perm.	Gene perm.	*	Sample perm.	Sample perm.	F-distribution

SC = self-contained, C = competitive, ORA = over-representation analysis, FCS = functional class scoring, M = multivariable, U = univariable, DAG = directed acyclic graph, UG = undirected graph, DG = directed graph, PCI = Pathway Connectivity Index, GGM = Graphical Gaussian Models, IPS = Iterative Proportional Scaling, GL = Graph Laplacian, * = various statistics are possible, for detection of differentially expressed pathways between two conditions authors suggests Mann-Whitney test.

https://doi.org/10.1371/journal.pone.0191154.t002

package (SPIA, TopologyGSA, Clipper, DEGraph) [18] and (iii) `ToPASeq` package [19] (all methods). ToPASeq package is our previous work in which we either de novo implemented or optimised existing implementations of a number of existing topology-based pathway analysis methods. For the sake of access uniformity for method application and access to method-specific pre-processing, we chose to use the `ToPASeq` package in our work.

The following section describes gene expression data matrices and pathway topologies used in each experiment. We do not define the basic terms from graph theory, since they are explained in many textbooks, for example [22]. Statistical details of individual experiments and key properties of the compared methods are described in the S1 Text.

Real datasets

In our study we used real gene expression microarray datasets from three public collections: Gene Overexpression Data Collection [23, 24], Breast Cancer Data Collection [25] and Disease Control Data Collection [26, 27]. These collections were obtained and pre-processed as described in the S1 Text. For each real dataset, we can anticipate one or several pathways that are expected to be differentially expressed or their identification is of particular interest due to experimental design. However, those pathways cannot be called 'true positive'. The Gene Overexpression Data Collection was selected because it allows us to study the effect of one perturbed gene. The Breast Cancer Data Collection represents a collection of datasets related to the same biological problem, and we focus on the reproducibility of the results. In the Disease-Control Data Collection, datasets cover various biological problems (cancer, metabolic, neurodegenerative diseases etc.) in a unified experimental design in which expression profiles of patients are compared to healthy controls. Additionally, we can identify a single pathway (*target pathway*) which is directly related to the particular disease and hence very likely to be differentially expressed. These datasets were used in Experiments 1, 6, 7 and 8.

Simulated datasets

Since the proper statistical distribution of the pathway expression data is unknown, we decided to use a real dataset (a dataset from Breast Cancer Data Collection denoted as VDX) as a base for the generation of simulated data. It contains 344 expression profiles of breast tumours obtained on an Affymetrix Human Genome U133A Array platform with 22 283 probesets corresponding to 13 091 unique Entrez IDs. We used estrogen receptor status as the main parameter dividing samples into two clinical groups. The simulated datasets were used in all experiments. The datasets for particular experiments were generated as shown in S1 Text.

Pathways and their topologies

We used human pathways from the KEGG database as the source of pathway topologies. For our comparison we used `graphite`'s pre-processed pathways as a default set of pathway topologies for the following reasons: (i) they are claimed to be superior to original implementation [28]; (ii) they allowed us to compare only the methods' algorithms regardless of the pre-processing strategy; (iii) the details of the pre-processing strategy are rarely described in the corresponding publication; (iv) the `graphite` implementation is readily available and widely used. In `ToPASeq` one can choose either `graphite` pre-processed pathways (+GPT) or pathway pre-processing as in the original implementation (MSPT) (if available) and hence evaluate the effect of different pre-processing strategies. The +GPT topologies were used in all experiments, and the MSPT was used in Experiment 8 only. In Experiment 7 we also used non-topological variants of the compared methods corresponding to pathway topologies without interactions (-GPT). To reduce computational complexity, we filtered out pathways with more than 150 genes and with less than two genes with available expression data.

Results

Experiment 1: Effect of sample size, pathway size, platform density and number of differentially expressed genes

Fig 2 shows the influence of sample size on the proportion of DEPs in both real and simulated data. In the simulated datasets (Fig 2A), an increase in sample size results in an increase in the proportion of DEPs for TAPPA and all the multivariable methods (TopologyGSA, Clipper, DEGraph). For each of these methods, we observed a breakpoint (sample size) beyond which the proportion of DEPs stabilised. For TopologyGSA and Clipper, this breakpoint was at 68 samples, with 94.9% and 93.4% DEPs, respectively. For the complete dataset (344 samples), DEGraph and TAPPA identified 94.2% and 68.7% of pathways to be differentially expressed, respectively. On the other hand, SPIA, PRS and CePa reported a rather stable proportion of differentially expressed pathways across all sample sizes (medians between 4.7% and 14.2%). Interestingly, there is a trend of decreasing number of DEPs with increasing sample size in CePa.

Similar observations were made in the analysis of real datasets from the three real data collections. Across all data collections, the highest proportion of DEPs was observed in Clipper (median: 92.5%), followed by TopologyGSA (median: 73.7%), DEGraph (median: 48.0%) and TAPPA (median: 36.1%). CePa, SPIA and PRS reported the smallest median proportion of DEPs (27.9%, 16.5% and 13.9%, respectively). Results for individual disease collections are shown in (Fig 2B). Although the Gene Overexpression Data Collection comprised of relatively small datasets (Table 3), multivariable methods still reported a large proportion of DEPs, similar to the case of generally larger datasets in the Breast Cancer Data Collection. The smallest dataset (with overexpressed c-Src) had the lowest proportion of DEPs in all methods.

The Breast Cancer Data Collection contained datasets of various microarray platform sizes (probes representing between 2 780 and 20 389 unique Entrez IDs). For *competitive* methods (SPIA, PRS and CePa), the statistical significance of the differential expression of a pathway depends on the set of genes measured in the experiment. A smaller number of genes outside a

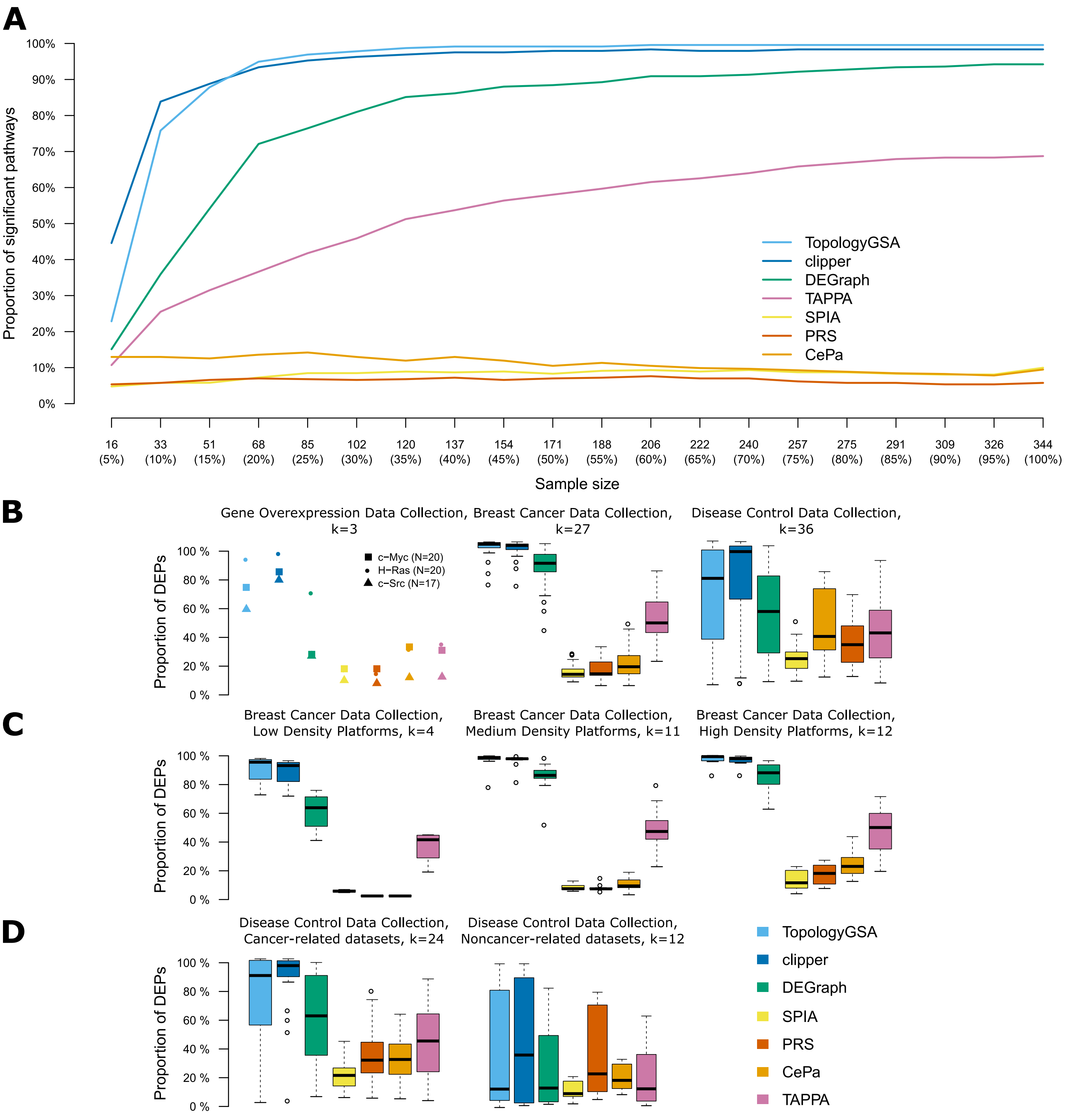


Fig 2. The effect of sample size. (A) The selected dataset from Breast Cancer Data Collection (denoted as VDX) was reduced to 20 random subsets representing 5%, 10%, . . . 95% of its original sample size (while preserving the proportion of samples in the clinical groups) leading to sample sizes from 16 to 326. Differentially expressed pathways between estrogen receptor positive and negative samples were detected. The lines show the median proportion of significant pathways ($p < 0.05$) over 20 subsets for each sample size. (B-D) Graphs indicating the percentage of differentially expressed pathways (DEPs) in the respective data collections. k denotes the number of datasets. See Table 3 for the summary of sample sizes. The datasets from the Breast Cancer Data Collection were divided by platfrom densities into: low-density platforms (2780-5486 EntrezIDs), medium-density platforms (9041-13091 EntezIDs) and high-density platforms (17779-20389 EntrezIDs).

https://doi.org/10.1371/journal.pone.0191154.g002

Table 3. Overview of the data collections.

Data Collection	Number of datasets	Sample size			Number of gene IDs		
		Median	Min	Max	Median	Min	Max
Gene Overexpression	3	17	17	20	23 521	23 521	23 521
Breast Cancer	27	129	49	856	13 091	2 780	20 389
Disease-Control	36	21	8	153	20 535	12 438	20 535

https://doi.org/10.1371/journal.pone.0191154.t003

pathway leads to reduced variability of the random sets of DEGs which results in lower probability of extreme pathway-statistic and, as consequence, higher p-value. Hence, we split the collection into low-, medium- and high- density platform datasets, based on the number of unique EntrezIDs their probes mapped to (from 2780 to 5486 EntrezIDs for low-density, 9041 to 13091 EntrezIDs for medium-density and 17779 to 20389 EntrezIDs for high-density platforms) (Fig 2C and S1 Fig). Indeed, all the competitive methods reported fewer DEPs in the datasets from low-density platforms. On the other hand, one *self-contained* method—DEGraph also reported fewer DEPs. In DEGraph, each pathway is divided into connected components which contain only the measured genes. In case of low-density microarray platform, this results in the small size of the individual components which tend to have higher *p*–values.

The Disease-Control Data Collection contained small to medium size datasets in which patients with various diagnoses were compared to healthy controls. The proportion of DEPs varied greatly between datasets from this collection (Fig 2B). However, when we divided the datasets into cancer-related and non-cancer-related, all the methods reported more DEPs for the cancer-related datasets (Fig 2D). We hypothesised that this was a consequence of the larger number of differentially expressed genes (it is known that tumours have highly deregulated gene expression in comparison to healthy tissue). The proportion of DEPs as a function of the number of DEGs is shown in S2 Fig. Indeed, the percentage of DEPs depended on the number of DEGs in multivariable methods and TAPPA, but not in SPIA, CePa and PRS. Since in ORA methods (SPIA, PRS, CePa), fixed thresholds were used to identify DEGs, we assessed the effect of three thresholds ($p < 0.05$, $p < 0.01$ and $p < 0.001$) on the proportion of DEPs (S4 Fig). For stricter thresholds ($p < 0.01$ and $p < 0.001$), in all methods, the number of DEPs increased with increasing sample size, as one would expect based on statistical properties of hypothesis testing. For $p < 0.05$, however, this trend holds only until a breakpoint in sample size, which is method specific: between 85-120 samples in CePa, between 222-257 samples in PRS and between 257-291 samples in SPIA. After the breakpoint, the number of DEPs rapidly decreases for $p < 0.05$.

To study the effect of pathway size, we divided pathways into small (<35 nodes) and large (≥35 nodes) (following [29]). S3 Fig shows the median *p*-value of pathways within each group as a function of dataset sample size for individual methods. Large pathways achieved lower median *p*-values in comparison to small pathways, independently on the dataset sample size, except PRS. In PRS, we observed the opposite effect starting at 137 (40%) samples. In multivariable methods, median *p*-values decreased very rapidly with increasing sample size, dropping below 0.01 at 33 (10%) and 51 (15%) for Clipper and TopologyGSA.

Experiment 2: Type I error rate

For all methods, the observed type I error rate was close to the expected 5% threshold, except for CePa (12.8%), see Table 4 and S5 Fig.

Table 4. Type I error rates: For each method the number (N) and the proportion (%) of rejected hypotheses out of 1000 tested is shown.

Method	Rejected hypotheses N(%)
SPIA	30 (3.0%)
PRS	38 (3.8%)
Clipper	45 (4.5%)
TopologyGSA	47 (4.7%)
DEGraph	55 (5.5%)
TAPPA	57 (5.7%)
CePa	128 (12.8%)

https://doi.org/10.1371/journal.pone.0191154.t004

Experiment 3: Effect of mean expression, difference in expression and topology of a single gene

In this experiment, we studied the effect of group-specific increase of expression of single genes in three selected pathways (increments of 0.1 to 2 in log2 fold change with step size 0.1 in 200 simulated datasets). The influence of a gene was quantified as a proportion of identified differentially expressed pathways across all simulations and increments. For simplicity, we divided the gene influence into five categories: very low influence (0%-20% DEPs), low influence (20%-40% DEPs), medium influence (40%-60% DEPs), high influence (60%-80% DEPs) and very high influence (80%-100% DEPs) (S6 Fig), respectively.

An induced change in a single gene had a much stronger influence on the results of multivariable methods than on the results of univariable methods. The median proportion of DEPs (combined across all induced differences) for multivariable methods was 82.5% for TopologyGSA, 82.3% for Clipper and 42.7% for DEGraph, compared to 29.3% for PRS, 25.9% for CePa, 15.4% for TAPPA and 12.8% for SPIA.

We further examined the effect of relative change of gene expression between the groups, the effect of gene mean expression and the effect of gene topology in a pathway (S6 Fig).

Fig 3 shows the proportion of DEPs across all genes in the Non-small cell lung cancer pathway as a function of the induced change, for each method separately. TopologyGSA and Clipper were very sensitive to the increase in the induced log2 fold-change of a gene. The higher the fold change, the higher the proportion of DEPs. In fact, both methods marked 96% of the simulations as DEPs at log2FC = 1. In all the other methods, the effect of the increased induced change was less dramatic, although monotone, except CePa that reached its plateau at the induced change of 0.6 (28.3%).

The influence of gene topology was in agreement with methods' algorithms (S6 Fig). In TopologyGSA and Clipper, all the tested genes had a high or very high influence on the detection of DEPs, regardless of their topological properties (Table 5). The proportion of DEPs was instead correlated with mean expression of the individual genes. The mean expression had no significant effect on the proportion of DEPs in other methods. In DEGraph, the genes with the highest influence were those without incoming interactions (root nodes). In SPIA, the most influential genes had none or only neutral (e.g. binding) incoming interactions and many downstream genes. In PRS, most of the genes had low influence on the pathway detection, except for RIG-I-like receptor signalling pathway, which contained four genes with medium influence. One of these genes was a common subunit of two multiprotein complexes. We observed a correlation of the gene influence with the number of gene interactions in PRS and TAPPA (Table 5). Although the number of interactions is one of the centralities (see S1 Text,

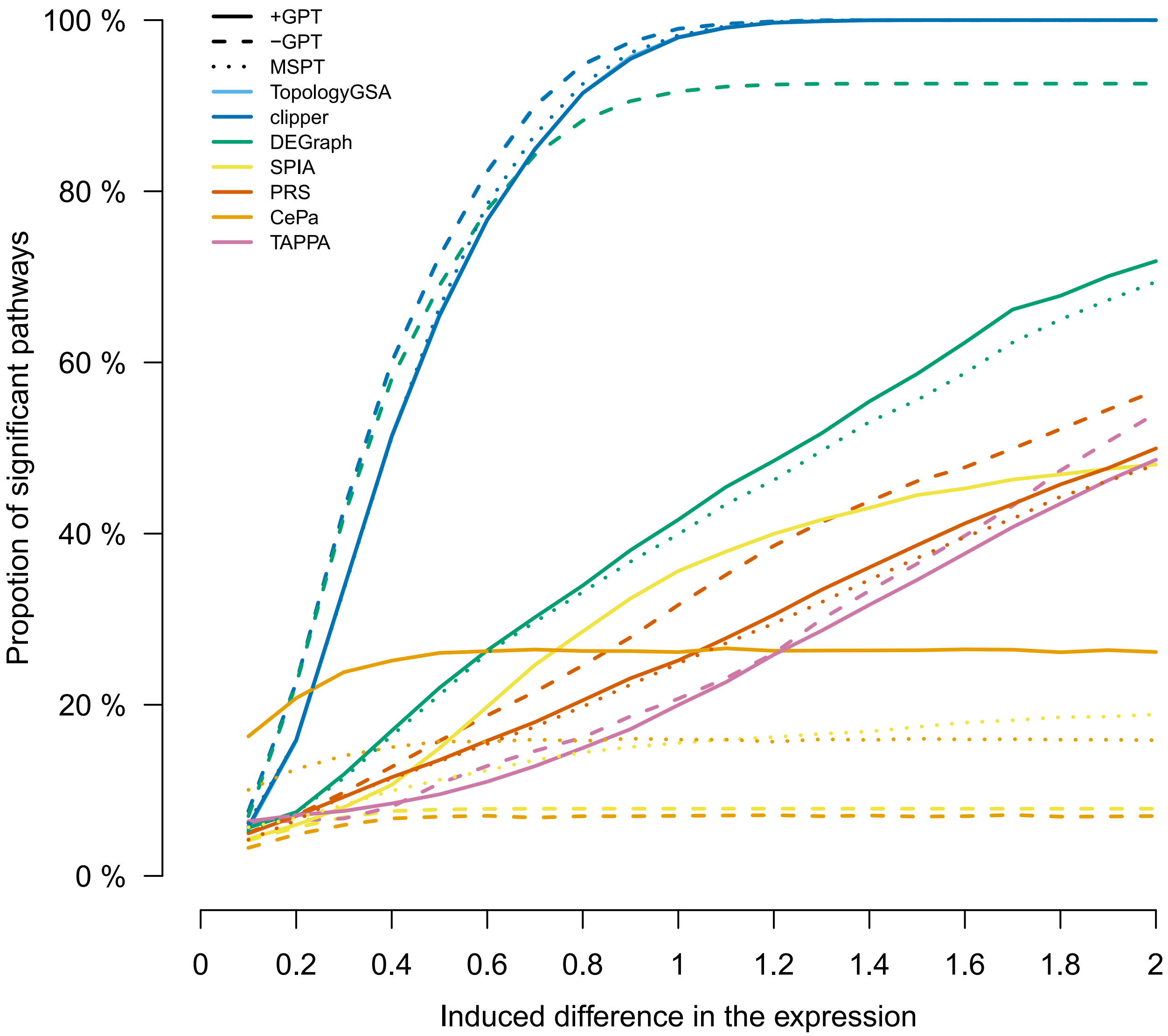


Fig 3. Proportion of differentially expressed pathways (DEPs) combined across all genes as function of the induced change. The proportion of differentially expressed pathways combined across all tested genes in the Non-small cell lung cancer pathway at different induced expression changes. Each line represents one method. Results were very similar for TopologyGSA and Clipper, and the respective lines are overlapping. Solid lines refer to pathway topology from graphite package (+GPT), dashed to pathway topology from graphite package without interactions (-GPT) and dotted to method-specific pathway topology (MSPT).

https://doi.org/10.1371/journal.pone.0191154.g003

section Materials and methods) used in CePa, the most influential genes were the nodes with the highest betweenness centrality.

Experiment 4: Effect of overexpression of multiple genes

Here, we assessed the combined impact of overexpression of multiple genes (gene sets), regardless of the possible topological motif. In all methods, the number of DEGs in a pathway positively correlated with the number of DEPs. Within the same gene set size, the influence of a gene set increased with the cumulative effect of individual genes as measured in Experiment 3 (S7 Fig).

Table 5. Spearman's correlations coefficients between the gene influence and the number of interactions stratified by interaction type.

	Pathway								
	Bacterial invastion of epithelial cells			Non-small cell lung cancer			RIG-I-like receptor signaling pathway		
	Interaction type			Interaction type			Interaction type		
Method	Incoming	Outgoing	Both	Incoming	Outgoing	Both	Incoming	Outgoing	Both
TopologyGSA	0.434	0.149	0.368	0.102	-0.123	-0.005	0.413	-0.063	0.239
Clipper	0.437	0.153	0.374	0.103	-0.120	-0.002	0.414	-0.059	0.244
DEGraph	-0.399	0.145	-0.264	-0.608	-0.158	-0.446	-0.585	-0.113	-0.477
SPIA	-0.153	0.278	0.096	-0.127	0.070	0.023	0.041	0.314	0.208
PRS	0.220	0.861	0.779	0.355	0.826	0.779	0.610	0.713	0.917
CePa	0.325	0.394	0.648	0.373	0.207	0.493	0.563	0.782	0.916
TAPPA	0.161	0.273	0.403	0.543	0.536	0.747	0.653	0.584	0.873

https://doi.org/10.1371/journal.pone.0191154.t005

Experiment 5: Effect of overexpression of topological motifs

In this experiment, we overexpressed three, four and five genes, respectively, representing one of the 18 topological motifs present in the Non-small cell lung cancer pathway (see S1 Text). Similarly to the previous experiments, the proportion of DEPs increased with the induced change and with the number of genes in the motif.

For the multivariable methods, we did not observe the influence of the motif on the proportion of DEPs when compared to the gene set effect from Experiment 4 (Fig 4). In all univariable methods, except SPIA, motif overexpression resulted in the increased proportion of DEPs in comparison to gene set overexpression. This difference in overexpression was independent of the number of overexpressed genes for TAPPA but diminished with the increasing number of overexpressed genes in PRS and CePa. In contrast, motif overexpression resulted in the decreased proportion of DEPs in SPIA in comparison to gene set overexpression.

The effect of the motifs in the context of previous findings and the motifs' properties (size, topology, the sum of effects of individual genes) is shown as a heatmap with information from Experiment 3 overlaid (S8 Fig, Fig 5). The heatmap shows clustered proportions of DEPs at different increments of log2 fold-changes (rows) in all tested topological motifs (columns). The proportion of DEPs increased with the induced change, and this effect separated the analysed motifs into multiple clusters. We categorised the motifs based on their overall effect (the proportion of DEPs from all the simulations and induced changes). We were also further interested to see how the clusters correlated with the size (3, 4 or 5 genes) and the topology of the motif. For all methods but TopologyGSA and Clipper, we observed a clustering of the motifs according to motif size (S8 Fig). Since Experiment 4 showed that effect of multiple genes is directly dependent on the sum of effects of individual genes, we plotted the effect of individual genes (as measured in Experiment 3) involved in the individual topological motifs in the panel below the heatmap. Here, the gene-specific influence is indicated by color (white means gene was not present in the motif). Clearly, in all methods, the impact of topological motif positively correlated with the impact of individual genes of the motif (S8 Fig).

Experiment 6: Identification of target pathways

In this experiment, for each real dataset we identified a pathway that was related to the disease or a biological problem and, in an ideal situation, this pathway should be detected as differentially expressed with very low *p*-value in comparison to other pathways.

During the analysis, we encountered multiple method-specific problems that resulted in the impossibility to analyse all available pathways. First, TopologyGSA requires the dataset to have

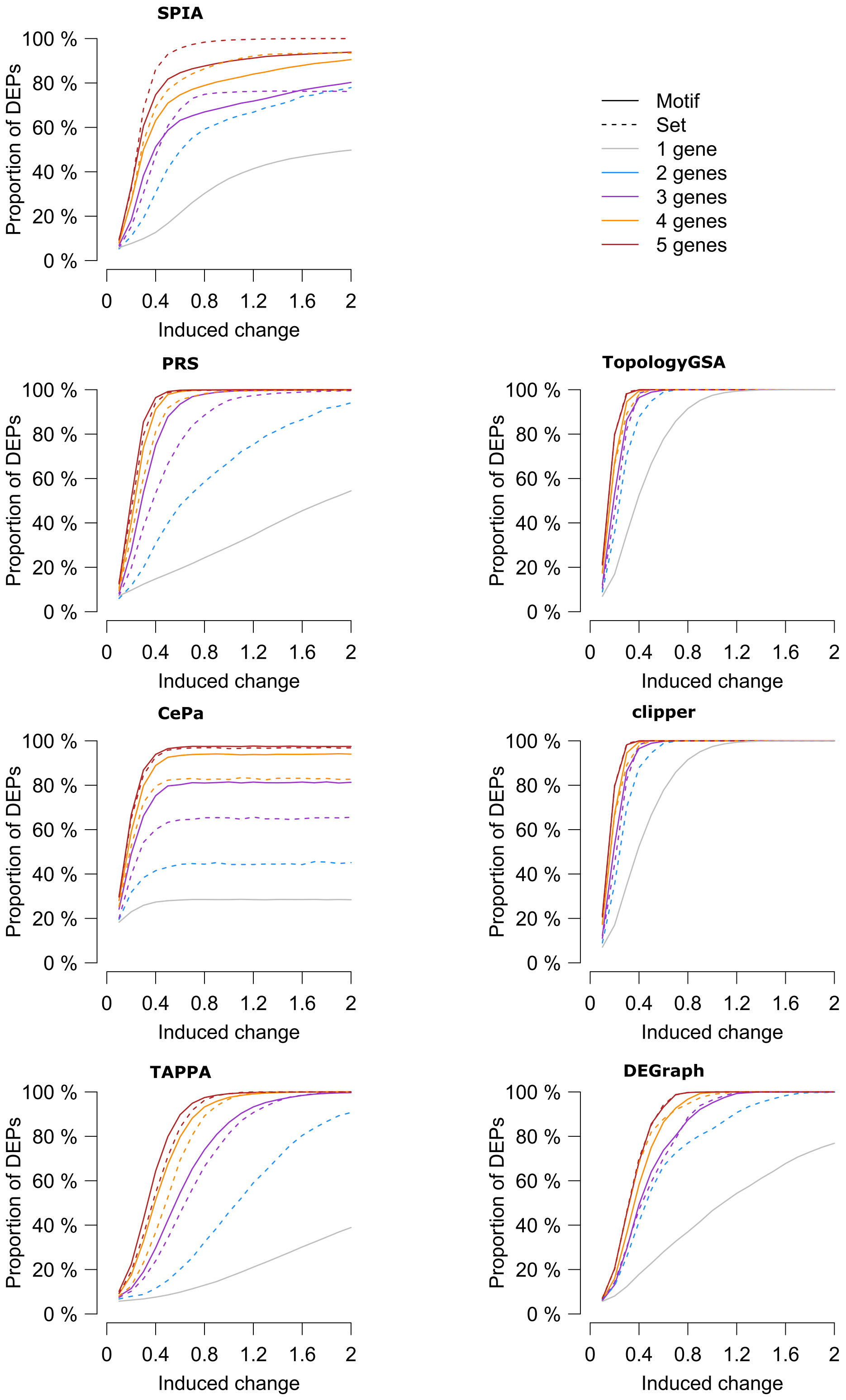


Fig 4. Comparison of the effect of expression change in a single gene, multiple genes and topological motifs. Combined influence of single gene, multiple genes and topological motifs on the proportion of differentially expressed pathways (DEPs) at varying induced expression changes is displayed. Sets of multiple genes and topological motifs are shown in the dashed and solid lines of the same color, respectively.

https://doi.org/10.1371/journal.pone.0191154.g004

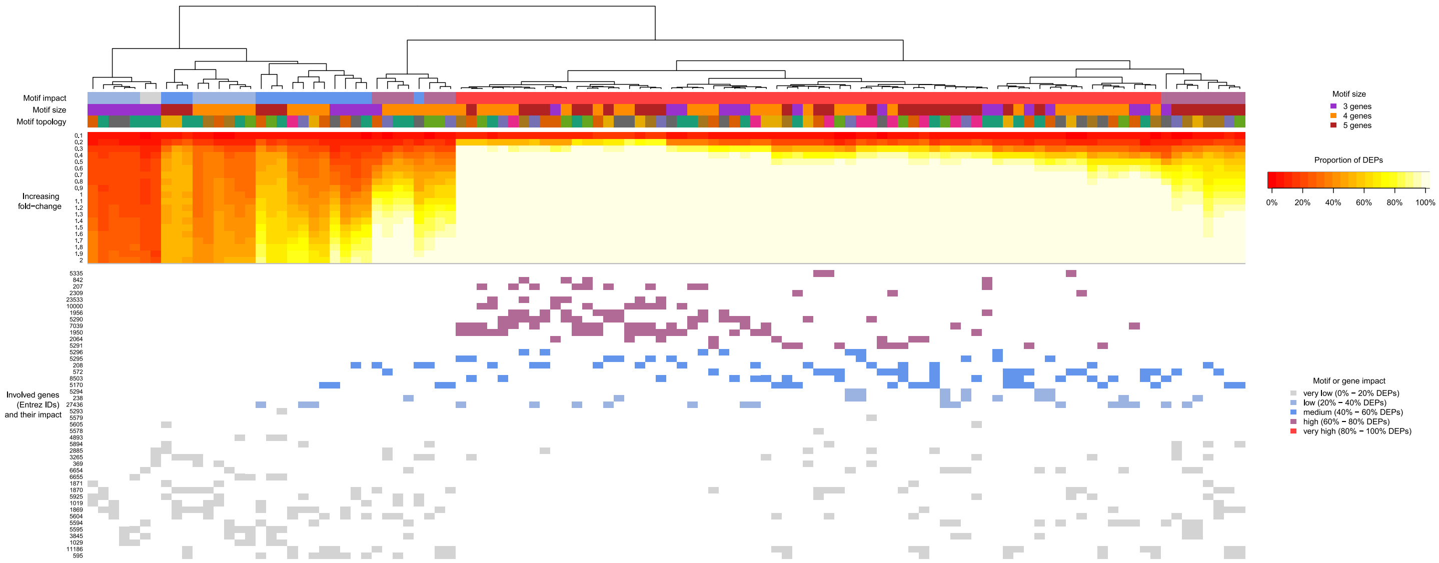


Fig 5. Effect of topological motifs in SPIA. Proportions of differentially expressed pathways (DEPs) for individual motifs (columns) at variable induced log2 fold-changes (rows) are displayed as a heatmap. Color bars on the top show influence of the motif, its size and topology (see S1 Text for details). Note, that colors used for motif topology are unique only among motifs of the same size. The bottom panel shows the influence of the genes in a representation of a topological motif as discovered in Experiment 3.

https://doi.org/10.1371/journal.pone.0191154.g005

more samples than the number of genes in the largest clique of the pathway and this condition was met only by several pathways. For large datasets, such as SUPERTAM_HGU133A from the Breast Cancer Data Collection (N = 856 expression profiles), we were unable to run TopologyGSA on 80GB RAM machine. DEGraph encountered similar but less frequent problems due to the singularity of pooled covariance matrices.

S9 and S10 Figs, and Tables 4 and 5 in S1 Text show results of the target pathway p-values and ranks in the Disease-Control Data Collection and Breast Cancer Data Collection. The results from the Gene Overexpression Data Collection can be found in S11 Fig. Since target pathways are unique for each dataset from this collection, they were not suitable for trend estimation.

Overall, multivariable methods assigned lower p–values and ranks to the target pathways than univariable methods. In the Disease-Control Data Collection, the target pathway was tested by TopologyGSA in only ten out of 36 datasets, of which nine times it was reported as differentially expressed. In contrast, the ranks from the DEGraph method were the highest in multivariable methods and the second largest in all methods. PRS and CePa reported consistently low median p-values (0.031 and 0.034, respectively) and low median ranks (19.5 and 25.5, respectively). Amongst univariable methods, the highest median p-value and rank of target pathways were observed in SPIA and TAPPA. In the Breast Cancer Data Collection datasets, the aim was to detect differentially expressed pathways between the estrogen receptor positive (ER+) and estrogen receptor negative (ER-) group. The set of target pathways therefore comprised of four pathways with estrogen receptor genes: Endocrine and other factor-regulated calcium reabsorption, Estrogen signalling pathway, Prolactin signalling pathway and Thyroid hormone signalling pathway. Since estrogen receptor plays different roles in these pathways and therefore harbours different topological 'importance', results for individual pathways from topology-based pathway analysis may vary. For all these pathways, all multivariable methods (TopologyGSA, Clipper, DEGraph) again reported very low p-values and ranks. From the univariable methods, TAPPA returned the lowest median p-values (except Estrogen

Table 6. Proportion of significant target pathways in the Gene Overexpression Data Collection.

Overexpressed gene in the target pathway	Method						
	SPIA	PRS	CePa	TAPPA	TopologyGSA	Clipper	DEGraph
c-Myc	7/15 (46.7%)	8/15 (53.3%)	12/15 (80%)	10/15 (66.7%)	0/0	14/14 (100%)	3/9 (33.3%)
H-Ras	14/40 (35.0%)	6/40 (15.0%)	16/40 (40.0%)	9/40 (22.5%)	1/1 (100%)	39/39 (100%)	18/23 (78.3%)
c-Src	5/14 (35.7%)	3/14 (21.4%)	3/14 (21.4%)	3/14 (21.4%)	0/0	14/14 (100%)	3/6 (50%)

https://doi.org/10.1371/journal.pone.0191154.t006

signalling pathway) and the highest ranks (except Endocrine and other factor-regulated calcium reabsorption pathway). The lowest median *p*-values and ranks of all target pathways amongst remaining univariable methods were observed in CePa. SPIA reports lower *p*-values and ranks than PRS only for the Endocrine and other factor-regulated calcium reabsorption pathway. Estrogen receptor is one of the root nodes and has a medium influence on this pathway in SPIA (47% DEPs) and only low influence in PRS (22% DEPs). On the other hand, Prolactin signalling pathway is the least significant by SPIA, and the estrogen receptor is a leaf node in this pathway with very low influence (3.5% DEPs). In the original experiments of the Gene Overexpression Data Collection, an overexpression of three genes (c-Myc, H-Ras, c-Src) was induced experimentally via adenoviral infection. The fold change of the perturbed genes ranged from 2.38 to 5.29 (S1 Text). 15, 40 and 14 target pathways were identified, for c-Myc, H-Ras and c-Src, respectively. The results of the analysis of this collection are summarized in Table 6. TopologyGSA was able to analyse only the Bladder cancer pathway, which was detected as differentially expressed. Clipper identified all target pathways as differentially expressed. Results of DEGraph, PRS, CePa and TAPPA, varied greatly between the three sets of pathways, ranging from 21% to 80% target pathways as differentially expressed. All univariable methods reported a higher percentage of target pathways as differentially expressed in the dataset with deregulated c-Myc in comparison to other datasets. When individual target pathways were assessed separately, DEGraph and univariate methods agreed on differential expression of the most biologically relevant pathways (S11 Fig).

Experiment 7: Effect of the exclusion of topological information

To assess the effect of exclusion of topological information, we studied the effect of individual genes on the proportion of differentially expressed pathways in the simulated datasets. We hypothesised that, in the non-topological setting, individual genes influence the final result equally. We applied the non-topological variants of the methods on both simulated (from Experiment 3) and real (from Experiment 6) datasets and Non-small cell lung cancer pathway was used as a model pathway for simulated data. Then we quantified the effect of genes in simulated datasets and computed the corresponding *p*–values and ranks of target pathways. The results were compared to the results obtained in Experiment 3 and the Experiment 6 (Fig 3).

The effect of the individual genes in simulated data is shown in S12 Fig. In TopologyGSA and Clipper, no difference between the topological and non-topological variant of the method was found. In all other methods, we did observe, in agreement with our hypothesis, the equal redistribution of the effect of the genes across the pathway in the non-topological variant. For DEGraph and PRS, the non-topological variant resulted in an overall increase of the individual gene effects, while in CePa and SPIA, the individual effects of the genes diminished. In the Disease-Control Data Collection (S9 Fig), we observed increased *p*-values and ranks for target pathways in PRS and CePa and decreased *p*-values and ranks for DEGraph and SPIA. No effect of exclusion of topological information was found in TAPPA, TopologyGSA and Clipper. Note that the median *p*-value of the target pathway was below 0.2 in all methods regardless pathway topologies. In PRS, the median *p*-value raised from 0.031 in the topological variant to 0.055 in the non-topological variant. In the Breast Cancer Data Collection (S10 Fig), we observed the pathway-specific effect of the exclusion of pathway topologies in SPIA where *p*-values increased only in the pathway in which estrogen receptor is one of the root nodes (Endocrine and other factor-regulated calcium reabsorption) and decreased in other pathways. In all estrogen receptor containing pathways, we observed increased *p*-values in CePa and decreased in PRS. No difference was observed in multivariable methods.

Experiment 8: Effect of pre-processing of pathway topologies

To assess the effect of pre-processing of pathway topologies (methods' original pre-processing MSPT vs `graphite` pre-processing +GPT), we first compared effects of the individual genes in model pathways (Fig 6). The main differences between +GPT and MSPT were in the pre-processing of multisubunit protein complexes, gene families and interactions related to non-gene product nodes (e.g. small chemical compounds). These differences had a direct effect on individual genes by changing their properties or an indirect effect on the genes by altering the distribution of a particular property in a pathway. No difference in the effects of individual genes was observed in Clipper. In the DEGraph's original pathway topology (MSPT) there were no interactions between subunits of multiprotein complexes. These interactions were introduced in `graphite` (S1 Text, [28]) pathway topologies (+GPT). In consequence, the genes whose products were subunits of multiprotein complexes had a different effect in MSPT compared to +GPT (see Fig 6, RIG-I-like receptor signalling pathway and Non-small cell lung cancer pathway). There were no protein complexes in the Bacterial invasion of epithelial cells pathway, so the gene effects were the same. In PRS, we observed a clear difference in the effect of individual genes only in the Non-small cell lung cancer, where a group of six genes had approximately two times higher effect in MSPT compared to +GPT. In this pathway, two nodes involved each of these genes—either as a member of two different gene families or a single node and a member of a gene family. In MSPT of PRS, gene families were processed into combined nodes (S1 Text), hence possibly increasing the effect of genes present in multiple nodes. We observed complex differences in gene effects between +GPT and MSPT for CePa. In CePa's MSPT, gene families and protein complexes are pre-processed into combined nodes, thus decreasing their degree centralities (if they interacted with other families or complexes) or decreasing the total number of nodes in a pathway resulting in the reduced influence of family members or subunits of protein complexes. At the same time, both the influence and the degree centrality of the genes interacting with these families was reduced. However, other genes gained importance as consequence of the different distribution of centralities or pathway topology. SPIA-specific pre-processing of pathway topologies did not propagate perturbations of individual genes through as many interaction types (including compound-mediated interactions) as in `graphite`. Therefore, in MSPT, the number of genes with high influence was reduced.

In both the Breast Cancer Data Collection and Disease Control Data Collection, with the agreement to the individual gene overexpression experiment, we observed increased *p*-values in CePa; slightly increased ranks in DEGraph and decreased *p*–values in PRS and no difference in *p*-values in Clipper (S10 Fig). For SPIA, we observed no difference in both *p*–values and ranks in agreement with the individual gene overexpression experiment only in the Breast Cancer Data Collection and decreased *p*–values and ranks in the Disease Control Data Collection.

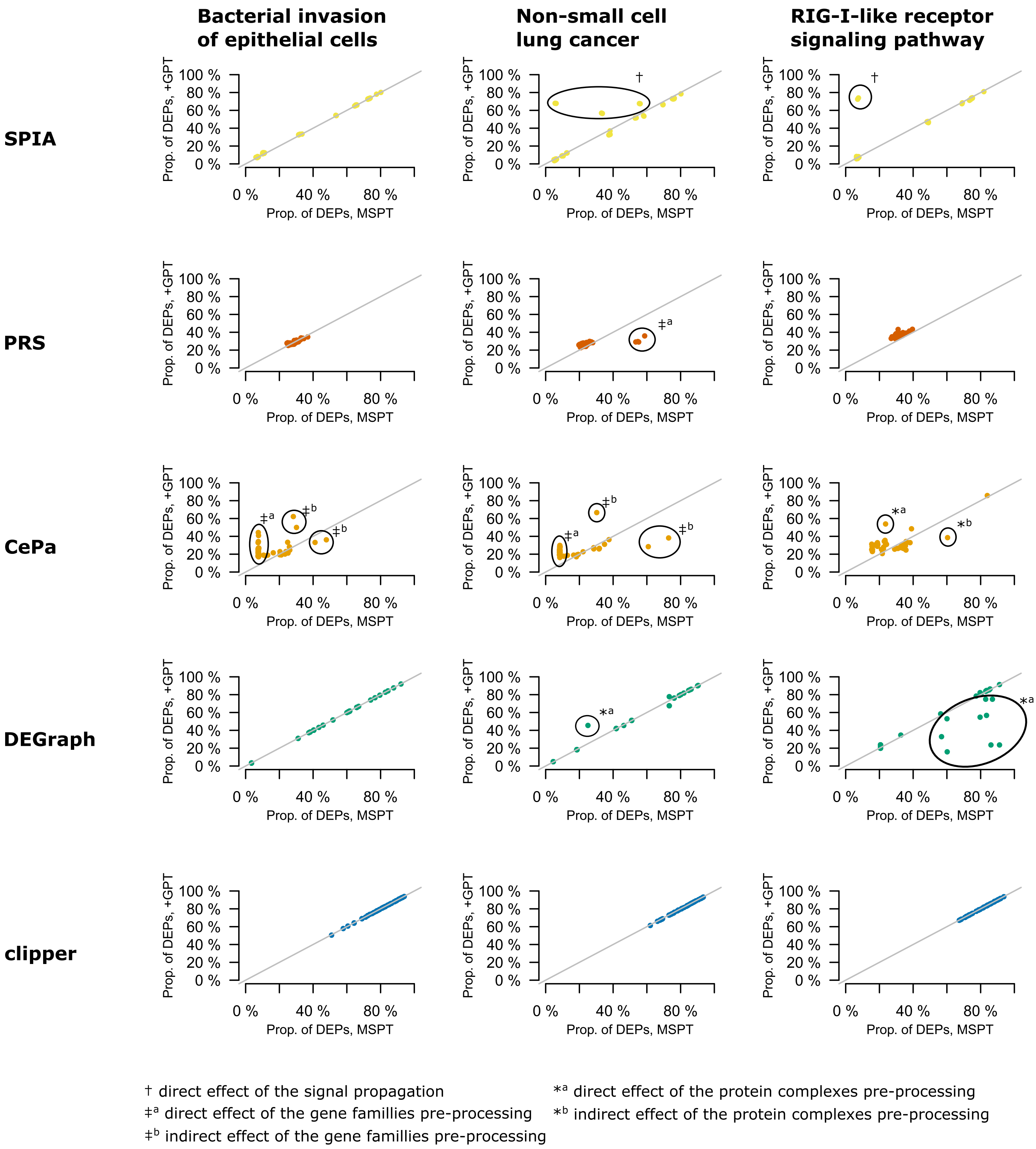


Fig 6. Effect of pre-processing of pathway topologies on simulated data—Overexpression of single gene. Each point represents a single gene. Only genes common for pathway topologies from graphite package (+GPT) and method-specific pathway topologies (MSPT) are displayed. Points on diagonal represent genes with the same influence in +GPT and MSPT. Points below (above) diagonal represent genes with higher (lower) influence in MSPT.

https://doi.org/10.1371/journal.pone.0191154.g006

Table 7. Overall assessment of the compared methods.

Parameter	SPIA	PRS	CePa	TAPPA	TopologyGSA	Clipper	DEGraph
Median proportion of DEPs in real datasets	16.5%	13.9%	27.9%	36.1%	73.7%	92.5%	48.0%
Effect on proportion of DEPs due to							
Increasing sample size	→	→	slowly ↘	↗	rapidly ↗	rapidly ↗	↗
Increasing pathway size	↑	↑↓	↑	↑	↑	↑	↑
DEGs threshold $p < 0.001$	↑	↑	↑	NA	NA	NA	NA
DEGs threshold $p < 0.01$	↑	↑	↑	NA	NA	NA	NA
DEGs threshold $p < 0.05$	↑↓	↑↓	↑↓	NA	NA	NA	NA
Single DEG [% DEPs]	12.8%	29.3%	25.9%	15.4%	82.5%	82.3%	42.7%
Characteristics of the most influential genes							
Crutial node property	root node	connected DEGs	betweenness	degree	mean expression	mean expression	root node
Incoming interactions	!!	-	!!	!!	-	-	!!
Outgoing interactions	!!	!!	!!	!!	-	-	-
Mean expression	-	-	-	-	!!	!!	-
Impacts of individual genes as observed on simulated data							
+GPT [% DEPs]	4.2 - 82.6	23.4 - 44.0	16.9 - 86.4	3.5 - 71.2	51.2 - 94.5	51.2 - 94.5	4.0 - 92.6
-GPT vs. +GPT	↓	↑	↓	↕	↑	↑	↑
MSPT vs. +GPT	↓	↑	↕	NA	NA	→	↑
Preferred scenario for hypotheses generation							
Number of DEGs	Many	Many	Many	Any	Few	Few	Few
Sample size	Any	Any	Any	Any	Small	Small	Small
Pathway of interest	Any	Any	Any	Any	Small	Small	Small
Experiment scale	Genome	Genome	Genome	Any	Any	Any	Any

→ - stable, ↘ - decrease, ↗ - increase, ↑ - higher, more, ↓ - lower, less, ↑↓ - trend changes at certain point, NA - not applicable, root node - node without incoming interactions, !! - important, - - not important, ↕ - both effects observed

https://doi.org/10.1371/journal.pone.0191154.t007

Discussion

We presented a series of eight controlled experiments designed to gauge the suitability of a number of topological pathway analysis methods to various analytical scenarios. Since topological information can be used in different ways and for different goals, in our study, we decided to focus on methods that (i) aim to detect differentially expressed pathways between two groups of interest, (ii) use a priori known pathway structures (topologies) and (iii) model each pathway separately. We described the performance of the selected methods on both simulated and real datasets.

We studied the methods' behavior from several perspectives: the sample size, pathway size, platform density, effect size, number of differentially expressed genes, gene topologies, platform density, gene sets and their topological motifs, the inclusion of topology information in the method's algorithm and different strategies for pre-processing of pathway topologies. The influence of the tested variables was assessed by comparison of the proportion of differentially expressed pathways, their p-values and ranks.

Table 7 shows the overall evaluation of the compared methods and summarises the most important observations from our experiments.

In all the compared methods, large pathways (> 35 genes) were assigned lower p-values than small pathways. Also, as expected, when a pathway contained more differentially expressed genes it was more often detected as differentially expressed. The number of differentially expressed genes usually surpassed their topological influence. None of the methods showed a preference for a particular differentially expressed topological motif.

The most striking difference was found between multivariable and univariable methods. Multivariable methods (TopologyGSA, Clipper and DEGraph) overall reported larger proportions of differentially expressed pathways in comparison to univariable methods (SPIA, PRS, CePa and TAPPA). Although all tested multivariable methods are derived from Hotelling's T^2 statistic, they differed significantly in their performance. TopologyGSA and Clipper assigned very low p-values and ranks to all the target pathways. However, this seems to be the result of overall low specificity, since they reported many other pathways (if not all) as differentially expressed. These methods were also sensitive to the increase in the sample and pathway size, the number of differentially expressed genes and the mean gene expression. The higher the increase, the lower the p–values and the larger the proportion of differentially expressed pathways, independent of the platform density. These findings indicate that in the scenario where (i) many differentially expressed genes are expected (e.g. cancer-related experiments); (ii) the dataset contains more than a few tens of samples ($>$ 68 samples in our experiments); (iii) a pathway contains a gene with at least a subtle random change in the expression, the pathway will be identified as significant. This behavior agrees with the *self-contained* nature of the methods, which is known to have higher sensitivity. However, many differentially expressed pathways identified by these methods might be false positives and therefore not useful for selection of biological hypotheses for further research. Interestingly, in TopologyGSA and Clipper, the exclusion of the topological information made no difference in the results. Therefore, despite well-established mathematical background (Graphical Gaussian models), these methods do not appear to fit the definition of topology-based methods for identification of differentially expressed pathways.

In contrast, DEGraph detected fewer differentially expressed pathways compared to TopologyGSA and Clipper, suggesting higher specificity. At the same time, in DEGraph the influence of individual genes was related to the pathway topology. DEGraph was less sensitive to sample size, pathway size or the number of differentially expressed genes. The performance of the non-topological variant of DEGraph was similar to the TopologyGSA and Clipper with or without topology. Different pathway pre-processing strategies had only limited influence on both DEGraph and Clipper (not assessed for TopologyGSA).

Univariable ORA methods SPIA, PRS and CePa, assigned low p-values only to some of the target pathways depending on the topological properties of differentially expressed genes in the pathway. This behaviour suggests higher specificity and stronger dependency on the topological information. These methods were less sensitive to the effects of sample size, pathway size, number of DEGs or thresholds used to identify differentially expressed genes. However, with increasing number of differentially expressed genes in a pathway, the effect of gene topology became less important. Due to the competitive nature of SPIA, PRS and CePa, these methods reported less differentially expressed pathways on low-density platforms. The univariable methods also exhibited higher sensitivity to the pre-processing of pathway topologies. Hence they can be considered true representatives of the topology-based pathway methods. Pre-processing of protein complexes, gene families and interactions involving non-gene products (metabolites such as PIP3) was the key factor in methods' performance and influence of the individual genes. Although, our results suggest that, for PRS and CePa, the method-specific pathway pre-processing seams to be more appropriate and should be preferred to `graphite`'s approach, further research is needed to identify an optimal pre-processing strategy for the compared methods. For instance, gene family members may be incomplete, and thus the observed increased influence of a gene which is a member of two different gene families may not be biologically sustained. Also, members of a gene family are seen as interchangeable regarding signal transduction, while each subunit of a protein complex is necessary for complex assembly and biological function. Therefore the unified approach, as used in method-specific pathway pre-processing, may not be optimal. The TAPPA [11] method stands out with its unique algorithm—a gene expression profile is being transformed into a pathway-level expression profile. Pathway-expression profiles were then analysed with traditional statistical methods (e.g. Mann-Whitney test for identification of differentially expressed pathways between two groups). As a consequence, this method is suitable also for applications with a complex experimental design. The sensitivity and specificity of TAPPA seemed to be well balanced. Amongst univariable methods, it was the most sensitive to sample size and usually identified most of the differentially expressed pathways. However, the proportion of differentially expressed pathways was never as high as in TopologyGSA or Clipper. At the same time, the method performance depended on the topological properties of the deregulated genes.

Guidelines for method selection

The increased sensitivity of multivariable methods (mainly TopologyGSA and Clipper) makes them ideal candidates for pathway analysis of experiments, where subtle changes in expression or a small number of differentially expressed genes between two conditions are expected—e.g. as in the case of tumor samples which contain a significant proportion of non-tumoral tissue (such as supporting stroma), thus confounding and diminishing measured signal of the gene expression. Since multivariable methods do not use lists of differentially expressed genes based on pre-defined thresholds but work with a complete list of the tested genes, they can be applied even in cases where none or very few genes are significant after statistical testing (for instance due to small sample size). The results of these methods, however, must be taken with caution and the significance of a pathway of interest must be interpreted in the context of all the results to ensure it is not just a consequence of overall low specificity of the method. To control for low specificity of the result, we recommend using DEGraph.

Univariable methods are not sensitive to the sample size or the number of differentially expressed genes in the datasets. Their ability to identify particular pathway as differentially expressed is highly dependent on the topological properties of the deregulated genes, the inclusion of the topological information and the pre-processing of the pathway topologies. Univariable methods are recommended in most applications and especially when the biological hypothesis aims at a pathway where genes of certain topological properties (biological function) are expected to be affected (see below and Fig 7B). However, since SPIA, PRS and CePa are ORA methods, they require at least some differentially expressed genes, and their applicability on datasets with very subtle changes in gene expression can be limited (in contrast to multivariable methods). On the other hand, if the differentially expressed genes occupy in the pathway the "correct" topological positions, the topological properties of the methods help to categorize this pathway as significant despite a small overall number of differentially expressed genes in the pathway. TAPPA, in contrast, being the FCS method, is a good choice for applications with a limited number of differentially expressed genes overall. Since in TAPPA the most important genes are those with many interactions, pre-processing of gene families and protein complexes must be carefully considered as their expansion into individual members or subunits may unintentionally increase their effect.

Based on our results we propose some guidelines for optimal method selection based either on (i) design of the experiment (comparison type, input data type, the platform density, sample size, expected number of differentially expressed genes)—Fig 7A; or (ii) selected (preferred) deregulation type—Fig 7B. Note, that the presence of many differentially expressed genes in a pathway surpasses topological effect of individual genes. Fig 7C shows an example of the most

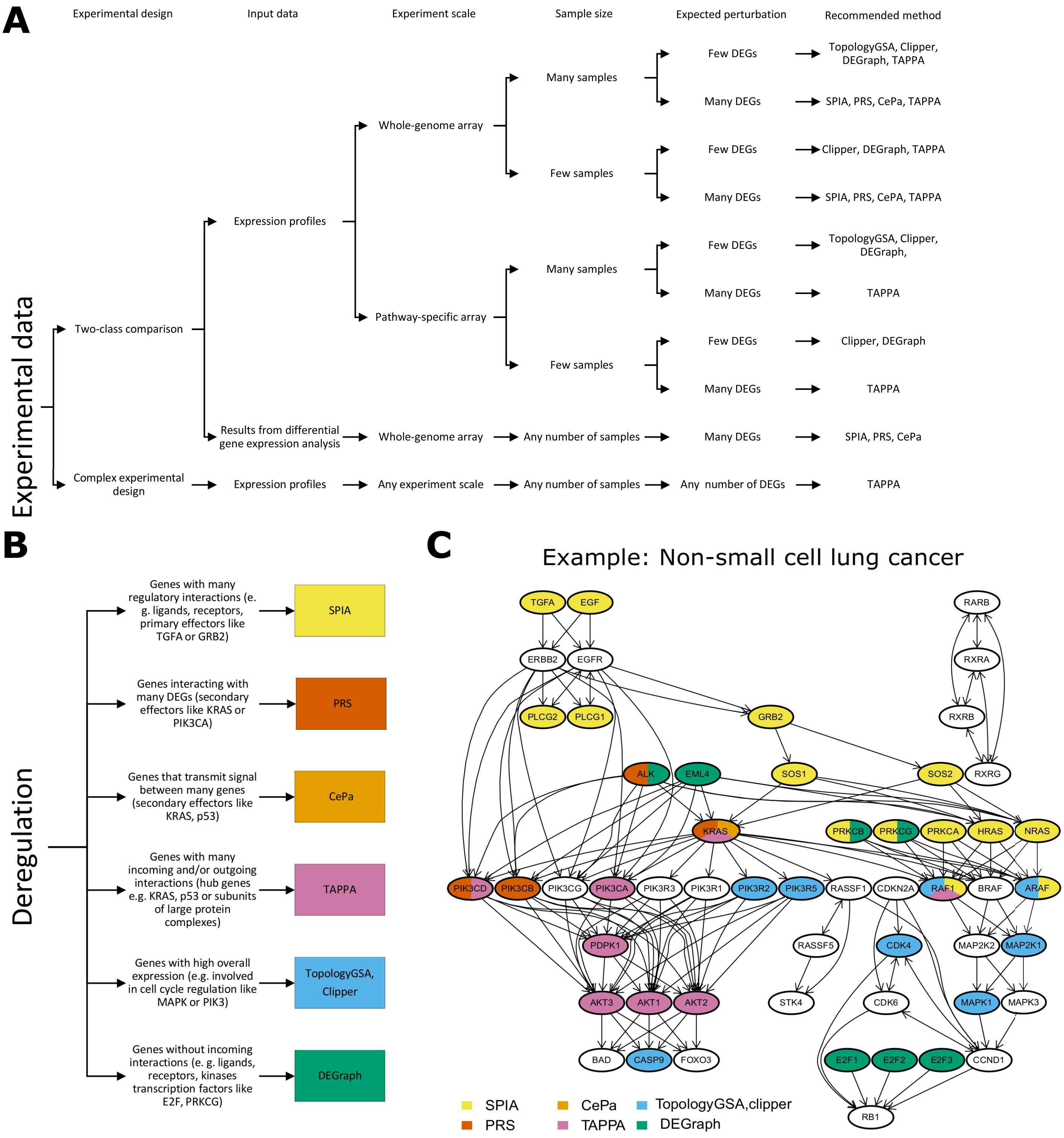


Fig 7. Guide to selection of topology-based pathway analysis method. (A) Recommended methods for specific scenarios based on experimental design, available input data, platform density, sample size and expected number of differentially expressed genes. (B) The most important deregulated genes in particular methods. Individual methods prefer different genes as the most important for pathway deregulation, and these preferences represent another factor for optimal method selection. The genes are defined mostly by their topological properties (e.g. number of interactions). Examples of genes must be interpreted within specific pathway (p53 signalling pathway for p53, Non-small cell lung cancer for others), and specific pathway pre-processing (graphite). (C) An illustrative example of the most important genes in the Non-small cell lung cancer pathway from KEGG database as available in the `graphite` package.

https://doi.org/10.1371/journal.pone.0191154.g007

influential genes in the Non-small cell lung cancer pathway based on `graphite` pre-processing of topologies (+GPT). In SPIA, CePa, TAPPA and DEGraph, we colored all the genes with the highest influence as defined in Experiment 3 for each method. Details of the topological as well as biological properties of the most influential genes are described in the S1 Text. In TopologyGSA and Clipper, the most influential genes have the highest overall expression (usually related to the cell cycle regulation [30]). In DEGraph and SPIA, the genes without incoming interactions have the largest impact. These genes are often represented by ligands, receptors, or transcription factors (E2F family dissociating from pRB). On the other hand, genes interacting with many other genes (e.g. secondary effectors, such as PIK3CA or KRAS) have the highest influence in PRS, CePa and TAPPA.

The observed differences between topological methods should be considered when the results of pathway analyses are to be compared across experiments in which different methods were used to detect differentially expressed pathways. Currently, SPIA is the most often cited method (282 citations from Web of Science Core Collection as of 14 February 2017). The other compared methods were mainly used in methodological publications, in which the general concepts were compared to the new method and only very rarely in applications.

Conclusion

We performed one of the largest studies of topology-based pathway analysis methods published to date. In this study, we compared seven methods that aim to detect differentially expressed pathways from expression data employing a priori known pathway topologies in their algorithm. The methods were ranked according to their sensitivity to sample and pathway size, ability to detect target pathways, the proportion of differentially expressed pathways, benefit from incorporating topological information and sensitivity to different pathway pre-processing strategies. We also verified type I error rates and described the influence of overexpression and topological properties of a single gene or gene sets on the detection of a pathway as differentially expressed by the selected methods.

We demonstrated that multivariable self-contained methods are very sensitive to the changes in gene expression within a pathway leading to the uninformative identification of over 90% pathways as differentially expressed. As a consequence, a significant result can be easily obtained for a particular pathway. On the other hand, univariable methods (mostly competitive) were less sensitive to subtle changes in gene expression but exhibited stable performance over a wide range of scenarios and benefited from the inclusion of topological information.

Finally, we proposed guidelines for method selection based on a number of variables connected to experimental design as well as biological hypotheses. Overall, we recommend any of the multivariable approaches to be used mainly for applications with small sample size and subtle changes in gene expression, whereas univariable methods should be preferred for genome-scale applications with large changes in gene expression. The pre-processing strategy for pathway topologies must be carefully considered for univariable methods, and further research is required to identify an optimal pre-processing strategy.

Supporting information

S1 Text. Details of the selected methods and used real data collection.
(PDF)

S1 Fig. Effect of the number of Entrez IDs. Proportion of DEPs depending on the number of Entrez IDs for datasets from Breast Cancer Data Collection. Each point represents one dataset.
(PDF)

S2 Fig. Effect of the number of DEGs. Proportion of DEPs depending on the number of DEGs for datasets from Disease-Control Data Collection. Each point represents one dataset.
(PDF)

S3 Fig. Effect of the pathway size.
(PDF)

S4 Fig. Effect of the thresholds used for DEG detection.
(PDF)

S5 Fig. Distribution of *p*-values from Experiment 2.
(PDF)

S6 Fig. Summarization of the Experiment 3. Dependence of the proportion of DEPs on the difference in expression induced between groups, the gene mean expression and its postion. In SPIA, neutral interactions were drawn in grey.
(PDF)

S7 Fig. Effect of expression change in randomly selected multiple genes on the proportion of differentially expressed pathways. Sets of 2, 3, 4 and 5 genes were randomly selected from Non-small cell lung cancer pathway. Each circle represents one of those sets. Expression of genes in the set was modified with increments of 0.1 to 2 with step size 0.1 in 200 simulated datasets. Border color indicates number of genes in the set. Vertical axis shows combined influence of the genes (proportion of differentially expressed pathways across all increments and datasets). Horizontal axis corresponds to sum of the influence of individual genes. Pie color (from grey to blue and red) represents the influence of a single gene (see Experiment 3 for details).
(PDF)

S8 Fig. Summarization of the Experiment 5. Heatmaps of the proportion of DEPs for all compared methods.
(ZIP)

S9 Fig. P-values and ranks of the target pathways—Disease-Control Data Collection details. Boxplots of *p*-values and rank of the estrogen receptor-containing pathways in Disease-Control Data Collection. Ranks are based on *p*-values. Pathway with the lowest *p*-value has rank 1. All pathways with the same *p*-value recieved same rank. The rank was incremented by one between subsequent *p*-values.
(PDF)

S10 Fig. P-values and ranks of the target pathways—Breast Cancer Data Collection details. Boxplots of *p*-values and rank of the estrogen receptor-containing pathways in Breast Cancer Data Collection. Ranks are based on *p*-values. Pathway with the lowest *p*-value has rank 1. All pathways with the same *p*-value recieved same rank. The rank was incremented by one between subsequent *p*-values.
(PDF)

S11 Fig. P-values of the target pathways—Gene Overexpression Data Collection details. Heatmaps of *p*-values of the overexpressed oncogene-containing pathways in Gene Overexpression Data Collection. Pathways are ordered by the number of methods in which they are differentially expressed ($p < 0.05$).
(PDF)

S12 Fig. Effect of individual genes in non-topological variants of the methods. (A) Proportion of differentially expressed pathways for different genes and the difference in expression

induced between groups. (B) Dependence of the proportion of differentially expressed pathways on the difference in the gene position. In the non-topological variants of the methods (-GPT) we observed reduced proportion of differentially expressed pathways and loss of its dependence on gene postion in all methods except TopologySGA an Clipper.
(PDF)

Author Contributions

Conceptualization: Ivana Ihnatova, Eva Budinska.

Data curation: Ivana Ihnatova.

Formal analysis: Ivana Ihnatova.

Funding acquisition: Vlad Popovici, Eva Budinska.

Investigation: Ivana Ihnatova, Eva Budinska.

Methodology: Ivana Ihnatova, Eva Budinska.

Project administration: Vlad Popovici, Eva Budinska.

Software: Ivana Ihnatova.

Supervision: Vlad Popovici, Eva Budinska.

Visualization: Ivana Ihnatova, Vlad Popovici, Eva Budinska.

Writing – original draft: Ivana Ihnatova.

Writing – review & editing: Ivana Ihnatova, Vlad Popovici, Eva Budinska.

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517
2. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012; 8(2):e1002375. https://doi.org/10.1371/journal.pcbi.1002375 PMID: 22383865
3. Emmert-Streib F, Tripathi S, Matos Simoes Rd. Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods. Biology Direct. 2012; 7(1):44. https://doi.org/10.1186/1745-6150-7-44 PMID: 23227854
4. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Frontiers in Physiology. 2015; 6:383. https://doi.org/10.3389/fphys.2015.00383 PMID: 26733877
5. Bayerlová M, Jung K, Kramer F, Klemm F, Bleckmann A, Beißbarth T. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinformatics. 2015; 16(1):334. https://doi.org/10.1186/s12859-015-0751-5 PMID: 26489510
6. Braun R, Shah S. Network Methods for Pathway Analysis of Genomic Data; 2015.
7. Gu Z, Liu J, Cao K, Zhang J, Wang J. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Systems Biology. 2012; 6(1):56. https://doi.org/10.1186/1752-0509-6-56 PMID: 22672776
8. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in Physiology. 2013; 4:278. https://doi.org/10.3389/fphys.2013.00278 PMID: 24133454
9. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim Js, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009; 25(1):75–82. https://doi.org/10.1093/bioinformatics/btn577 PMID: 18990722
10. Al-Haj Ibrahim M, Jassim S, Cawthorne MA, Langlands K. A Topology-Based Score for Pathway Enrichment. J Comput Biol. 2012;.

11. Gao S, Wang X. TAPPA: topological analysis of pathway phenotype association. Bioinformatics. 2007; 23(22):3100–3102. https://doi.org/10.1093/bioinformatics/btm460 PMID: 17890270
12. Massa M, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. BMC Systems Biology. 2010; 4(1):121. https://doi.org/10.1186/1752-0509-4-121 PMID: 20809931
13. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Research. 2012; https://doi.org/10.1093/nar/gks866 PMID: 23002139
14. Jacob L, Neuvial P, Dudoit S. Gains in Power from Structured Two-Sample Tests of Means on Graphs. ArXiv e-prints. 2010;.
15. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology. 2004; 3(1). https://doi.org/10.2202/1544-6115.1027 PMID: 16646809
16. R Core Team. R: A Language and Environment for Statistical Computing; 2014. Available from: http://www.R-project.org/.
17. Huber W, Carey J V, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015; 12(2):115–121. https://doi.org/10.1038/nmeth.3252 PMID: 25633503
18. Sales G, Calura E, Romualdi C. graphite: GRAPH Interaction from pathway Topological Environment; 2016.
19. Ihnatova I, Budinska E. ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. BMC Bioinformatics. 2015; 16(1):350. https://doi.org/10.1186/s12859-015-0763-1 PMID: 26514335
20. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Research. 2007; 17(10):000. https://doi.org/10.1101/gr.6202607
21. Khatri P, Draghici S, Tarca AL, Hassan SS, Romero R. A system biology approach for the steady-state analysis of gene signaling networks. In: Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications. CIARP'07. Berlin, Heidelberg: Springer-Verlag; 2007. p. 32–41. Available from: http://dl.acm.org/citation.cfm?id=1782914.1782919.
22. Junker BH, Schreiber F. Analysis of Biological Networks. Wiley Series in Bioinformatics. Wiley; 2011. Available from: https://books.google.cz/books?id=YeXLbClh1SIC.
23. Kim JW, Mori S, Nevins JR. Myc-Induced MicroRNAs Integrate Myc-Mediated Cell Proliferation and Cell Fate. Cancer Research. 2010; 70(12):4820–4828. https://doi.org/10.1158/0008-5472.CAN-10-0659 PMID: 20516112
24. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006; 439(7074):353–7. https://doi.org/10.1038/nature04296 PMID: 16273092
25. Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, et al. A three-gene model to robustly identify breast cancer molecular subtypes. Journal of the National Cancer Institute. 2012; 104(4):311–325. https://doi.org/10.1093/jnci/djr545 PMID: 22262870
26. Bhatti G, Tarca AL. KEGGdzPathwaysGEO: KEGG Disease Datasets from GEO; 2012.
27. Bhatti G. KEGGandMetacoreDzPathwaysGEO: Disease Datasets from GEO; 2014.
28. Sales G, Calura E, Cavalieri D, Romualdi C. graphite—a Bioconductor package to convert pathway topology to gene network. BMC Bioinformatics. 2012; 13(1):20. https://doi.org/10.1186/1471-2105-13-20 PMID: 22292714
29. Tripathi S, Emmert-Streib F. Assessment Method for a Power Analysis to Identify Differentially Expressed Pathways. PLOS ONE. 2012; 7(5):1–13. https://doi.org/10.1371/journal.pone.0037510
30. Karlin S, Mrázek J, Campbell A, Kaiser D. Characterizations of Highly Expressed Genes of Four Fast-Growing Bacteria. J Bacteriol. 2001; 183(17):5025–5040. https://doi.org/10.1128/JB.183.17.5025-5040.2001 PMID: 11489855

[5] Xie T, D' Ario G, Lamb JR, Martin E, Wang K, Tejpar S, Delorenzi M, Bosman FT, Roth AD, Yan P, Bougel S, Di Narzo AF, Popovici V, **Budinská E**, Mao M, Weinrich SL, Rejto PA, Hodgson JG. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. PLoS One. 2012;7(7):e42001. doi: 10.1371/journal.pone.0042001. Epub 2012 Jul 31. PMID: 22860045; PMCID: PMC3409212.

A Comprehensive Characterization of Genome-Wide Copy Number Aberrations in Colorectal Cancer Reveals Novel Oncogenes and Patterns of Alterations

Tao Xie[1]*, Giovanni d' Ario[2], John R. Lamb[1], Eric Martin[1], Kai Wang[1], Sabine Tejpar[3], Mauro Delorenzi[2,4], Fred T. Bosman[4], Arnaud D. Roth[5], Pu Yan[4], Stephanie Bougel[4], Antonio Fabio Di Narzo[2], Vlad Popovici[2], Eva Budinská[2], Mao Mao[1], Scott L. Weinrich[1], Paul A. Rejto[1], J. Graeme Hodgson[1]*

1 Oncology Research, Pfizer Worldwide Research and Development, San Diego, California, United States of America, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium, **4** Lausanne University Medical Center, Lausanne, Switzerland, **5** Geneva University Hospital, Geneva, Switzerland

Abstract

To develop a comprehensive overview of copy number aberrations (CNAs) in stage-II/III colorectal cancer (CRC), we characterized 302 tumors from the PETACC-3 clinical trial. Microsatellite-stable (MSS) samples (n = 269) had 66 minimal common CNA regions, with frequent gains on 20 q (72.5%), 7 (41.8%), 8 q (33.1%) and 13 q (51.0%) and losses on 18 (58.6%), 4 q (26%) and 21 q (21.6%). MSS tumors have significantly more CNAs than microsatellite-instable (MSI) tumors: within the MSI tumors a novel deletion of the tumor suppressor WWOX at 16 q23.1 was identified ($p<0.01$). Focal aberrations identified by the GISTIC method confirmed amplifications of oncogenes including EGFR, ERBB2, CCND1, MET, and MYC, and deletions of tumor suppressors including TP53, APC, and SMAD4, and gene expression was highly concordant with copy number aberration for these genes. Novel amplicons included putative oncogenes such as WNK1 and HNF4A, which also showed high concordance between copy number and expression. Survival analysis associated a specific patient segment featured by chromosome 20 q gains to an improved overall survival, which might be due to higher expression of genes such as EEF1B2 and PTK6. The CNA clustering also grouped tumors characterized by a poor prognosis BRAF-mutant-like signature derived from mRNA data from this cohort. We further revealed non-random correlation between CNAs among unlinked loci, including positive correlation between 20 q gain and 8 q gain, and 20 q gain and chromosome 18 loss, consistent with co-selection of these CNAs. These results reinforce the non-random nature of somatic CNAs in stage-II/III CRC and highlight loci and genes that may play an important role in driving the development and outcome of this disease.

Citation: Xie T, d' Ario G, Lamb JR, Martin E, Wang K, et al. (2012) A Comprehensive Characterization of Genome-Wide Copy Number Aberrations in Colorectal Cancer Reveals Novel Oncogenes and Patterns of Alterations. PLoS ONE 7(7): e42001. doi:10.1371/journal.pone.0042001

Editor: Alejandro H. Corvalan, Pontificia Universidad Catolica de Chile, Chile

Received April 25, 2012; **Accepted** June 28, 2012; **Published** July 31, 2012

Funding: This work was supported in part by a grant from the Swiss National Science Foundation (SNF 320030_135421) and Foundation Medic and a grant of the Krebsforschung Schweiz (KFS 02697-08-2010) to AR and MD. ST is a senior clinical investigator of the Fund for Scientific Research Flanders and has received research grants from the Belgian Federation Against Cancer and from the Belgian National Cancer Plan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: TX, JL, KW, MM, SW, PR and JGH are or were employed by Pfizer Inc. However, this does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials. The authors have declared that no other competing interests exist.

* E-mail: tao.xie@pfizer.com (TX); hodgson.graeme@gmail.com (JGH)

Introduction

Colorectal cancer (CRC) ranks second to lung cancer in both incidence and mortality in developed countries [1]. It is characterized by highly complex patterns of somatic genetic alterations of oncogenes and tumor suppressors that drive initiation and progression [2,3,4]. Understanding the cellular and molecular mechanisms by which these genetic changes facilitate colon cancer formation is critical for development of targeted therapeutic strategies aimed at controlling disease progression while minimizing toxic side effects.

One well-established genetic mechanism by which cancer cells alter the activity of oncogenes and tumor suppressors is through changes in gene dosage. Detailed characterization of DNA copy number aberrations (CNAs) have helped identify important oncogenes including ERBB2 and EGFR, as well as tumor suppressors such as TP53 [5]. Numerous studies have documented genome-wide somatic CNAs in CRC [6,7,8,9,10,11,12,13,14,15,16,17,18], some of which have been linked to clinical outcome or metastatic progression [19,20,21,22,23,24]. However, many of these studies have been limited by modest sample size, low resolution assays, or lack of associated clinical annotation, particularly for early-stage (II/III) colon cancer. Consequently, a comprehensive overview of CNAs and their association with outcome in stage II/III colon cancer has not been developed.

We surveyed somatic CNAs in a collection of 302 stage II/III colon cancers derived from the Pan-European Trials in Adjuvant Colon Cancer (PETACC)-3 trial, a large randomized phase III assessment of the role of irinotecan added to fluorouracil (FU)/leucovorin (FA) as adjuvant treatment for colon cancer [25]. The results presented herein explore the relationship between CNA, mRNA [26] and outcome, and contribute to a comprehensive molecular overview of stage-II/III colon cancer, which is paramount for refining patient classification and effective treatment.

Materials and Methods

Clinical and mRNA Data for PETACC-3 Patients

All stage II/III colon cancer patients included in this study were derived from the PETACC-3 clinical trial [25], with at least 5 years of clinical follow-up for each patient. The age, gender, stage, MSI (microsatellite-instable) as well as BRAF and KRAS mutation status of the patient population are listed in **Table S1**. mRNA expression data was generated on the ALMAC Colorectal Cancer DSA platform (Craigavon, Northern Ireland), as reported previously [26]. Patient and ethics approval for this study was obtained from the PETACC-3 Translational Research Working Party (PTRW).

Molecular Inversion Probe Data Generation

DNA extractions were performed on macrodissected formalin-fixed, paraffin-embedded (FFPE) tumor tissue derived from a single 5 uM slide from 835 patient samples. Tumor tissue within each section was identified and labeled by a qualified pathologist (F. Bosman). For normal controls, DNA was extracted from samples with sufficient amounts of histopathologically normal adjacent tissue well away from the tumor margins. DNA was quantified using the picogreen assay. For samples that yielded less than the recommended input DNA amount (75 ng), all DNA was carried forward into the Molecular Inversion Probe (MIP) amplification, labelling, and hybridization protocols using Affymetrix's OncoScan V1.0 FFPE Express services (Affymetrix, CA). Samples that failed PCR amplification or displayed a Median Average Pairwise Difference (MAPD) >0.6 after hybridization were removed from the final analysis, resulting in 302 tumor samples along with 44 adjacent normal samples as the normal baseline comparator. Typically samples below 20 ng of input DNA failed the MIP amplification cutoff and were not carried forward to array hybridization. Samples with at least 75 ng of input DNA universally yielded high quality copy number data (MAPD<0.6). Results varied for input DNA amounts of 20–75 ng, where the MAPD>0.6 filter served to eliminate excessively noisy samples.

Copy Number Data Analysis

Copy number data was analyzed with the Nexus Copy Number 6.0 software (Biodiscovery, Inc., CA, USA). The raw copy number data for each probe provided by Affymetrix was smoothed by a quadratic correction provided by NEXUS and centered using diploid regions. CNA frequency comparisons amongst sample groups (e.g. MSS versus MSI; stage-II versus stage-III) was performed using NEXUS default thresholds of >15% difference and significance p<0.01 (Fisher's exact test). To generate copy number segments and minimal common regions (MCRs), we applied a modified version of the Circular Binary Segmentation (CBS) algorithm [27] called "Rank Segmentation" in NEXUS. The p-value cutoff for CBS was 1.0E–6, and segments were assigned to 1 of 5 bins: amplified (>3.8 copies), gained (2.3 to 3.8 copies), unchanged (1.7 to 2.3 copies), deleted (0.5 to 1.7 copies) or homozygously deleted (<0.5 copies). For MCR frequency significance testing, we used a p-value cutoff of <0.01 from the statistical Significance Testing for Aberrant Copy number (STAC) method [28]. Hierarchical clustering of CNA was performed in NEXUS too (complete linkage, sex chromosomes ignored). To detect focal amplifications, we applied GISTIC (Genomic Identification of Significant Targets in Cancer) version 2.0 [29] using a Q-value cutoff <0.25. Genes reported in GISTIC2 amplification peaks were further examined if they are enriched in any biological pathways. We used canonical pathway database provided by MSigDB [30]. Pathway gene sets with less than 10 members or greater than 500 members were excluded. Fisher's exact test was used to access if those genes are over-represented. FDR was calculated based on 100 permutations where random sets of genes of same size were tested. We also used Fisher's exact test to see if frequencies of certain CNAs differ among patient groups (stage II vs. III, MSI vs. MSS etc). Survival analysis was performed using the Kaplan–Meier method with a p value (log-rank test) cutoff of <0.01. For analysis of CNA/CNA correlations, the Pearson correlation was computed at the gene level for all pairs of genes as described previously [31]. To derive gene level summaries from the copy number data, we assigned the copy number values from the segment(s) overlapping each gene: when there were multiple segments within the gene boundary, we averaged the copy numbers from those segments. All genome-based data reported in this manuscript are based on NCBI build 36 (hg18) of the human genome.

Expression Data Analysis

Gene expression data from the PETACC-3 patients was reported previously [26]. We matched it with gene level copy number data by ENTREZ ID. Copy number and gene expression data were simultaneously available for 213 of the 269 MSS patients with available CNA data. To test cis-correlation between a gene's copy number and its own mRNA expression level across tumors, we categorized patients according to their aberration status (amplification, gain, no-change, loss or homozygous deletion) associated to the expression values of probe sets mapping to the same gene.

Results

Copy Number Aberrations and Microsatellite Instability

33 of the 302 samples in our analysis were microsatellite instable (MSI): consistent with previous studies [19,32], the average number of CNAs in MSI tumors (10.2±6.5) was significantly smaller (p<0.01, two sample t-test) than the average number of CNAs in microsatellite stable (MSS) tumors (33.2±17.6). Nevertheless, two focal regions were deleted significantly more frequently in MSI samples: chr16q23.1 (chr16:77,231,391–77,261,567 bp) in 24.2% of MSI samples vs. 7.1% of MSS samples (p<0.01), and chr20q11.1 (chr20:28,118,678–28,244,164) in 24.4% of MSI samples vs. 8.9% in MSS samples (p<0.01). Interestingly, the only gene contained within the 16 q23.1 locus is the WWOX tumor suppressor, an inhibitor of the WNT/beta-catenin pathway [33], which is frequently activated in colon cancer.

Recurrent CNAs, Novel Oncogenes and Affected Pathways

Given the relatively low CNA prevalence in MSI tumors, we focused our analyses on the 269 MSS tumors. As has been reported previously [7,8,9,10,11,12,13,14,15,16,17], the frequencies of copy number gains and losses across the genome were not randomly distributed (**Figure 1A**), with CNAs ranging from single copy gains and losses of broad chromosomal regions, to focal homozygous deletions and high-level amplifications (**Figure 2**). The most frequent regions of gain encompassed chromosomal regions 7 p, 8 q, 13 q, and 20 q, and the most frequent regions of loss encompassed 8 p, 17 p, and 18 q (**Figure 1A**).

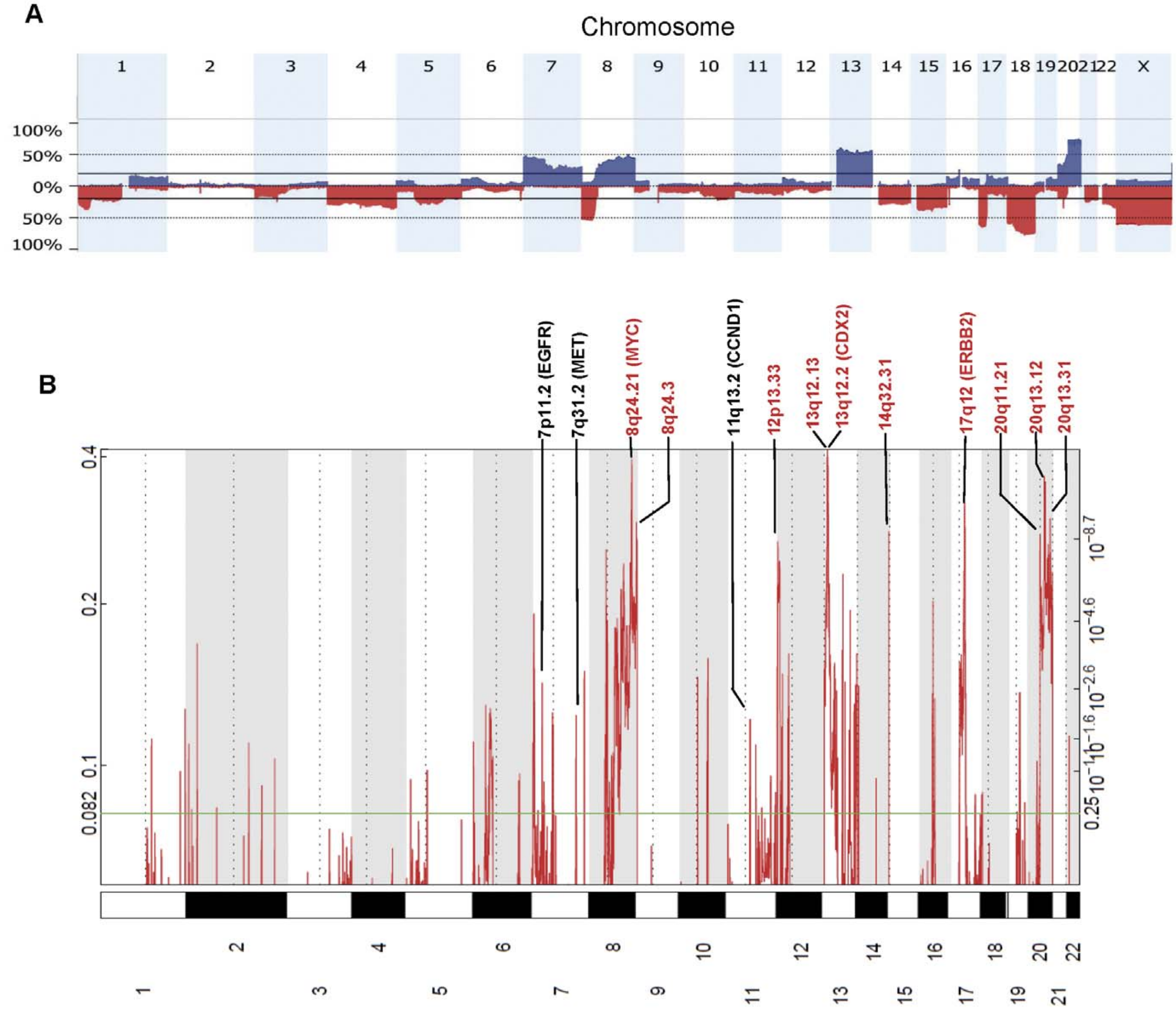


Figure 1. Summary of copy number aberrations detected in 269 MSS stage II/III colon cancer samples. (**A**) Frequencies of copy number gain (above axis, blue) and copy number loss (below axis, red) across the human genome. (**B**) Significance of focal amplifications detected by GISTIC 2.0. Chromosome positions were indicated along the y axis with centromere positions indicated by dotted lines. The ten most significant GISTIC peaks are shown in red text. Additional GISTIC peaks encoding established oncogenes are in black text. Details for all GISTIC peaks are provided in Table S3.
doi:10.1371/journal.pone.0042001.g001

To gain further insight, we summarized recurrent chromosomal gains and losses into Minimal Common Regions (MCRs) using Significant Testing of Aberrant Copy Number (STAC) [28], and GISTIC [29] to highlight candidate oncogenes in the MCRs based on the focality and amplitude of copy number change. A total of 66 MCRs were identified at frequencies above 10% (**Table S2**): there were 25 MCRs of gain ranging from 251 Kb to 104 Mb, and 41 MCRs of loss ranging from 286 kb to 138 Mb. GISTIC helped to refine the MCRs to loci and genes of particular significance (T**able S3).** Many of the significant peaks identified by GISTIC contained established oncogenes including CCND1, CDX2, EGFR, ERBB2, MET, and MYC (**Figure 1B**), along with tumor suppressors such as APC, SMAD4, and TP53. Several of the oncogenic peaks were driven by high-amplitude focal events in a subset of tumors (**Figure 2**), and these focal amplifications led to significant increases in mRNA expression for several of these genes. Highly significant GISTIC peaks not associated with well-established oncogenes or tumor suppressors include 12 p13.33 (**Figure 2E, F**) and 20 q13.12 (**Figure 2G, H**), which had recurrent high-magnitude focal amplifications, as well as 14 q32.31 which, although not highly amplified, had gains of sufficient recurrence and focality as to render a highly significant GISTIC Q-value (**Figure 1B**, **Table S3**). With the GISTIC amplicon data, we summarize 114 candidate cancer drivers in **Table S4**, which include twelve (10%) established oncogenes such as MYC, KRAS, and MET. Putative oncogenes including WNK1 (**Figure 3A**) and HNF4A (**Figure 3B**) have Q-score, amplified frequency, and cis-acting effects on mRNA that are comparable to established oncogenes (**Figure S1).** Our analysis has narrowed more than 6,000 genes from MCR regions of the genome to a manageable number of about 100 for further experimental validation.

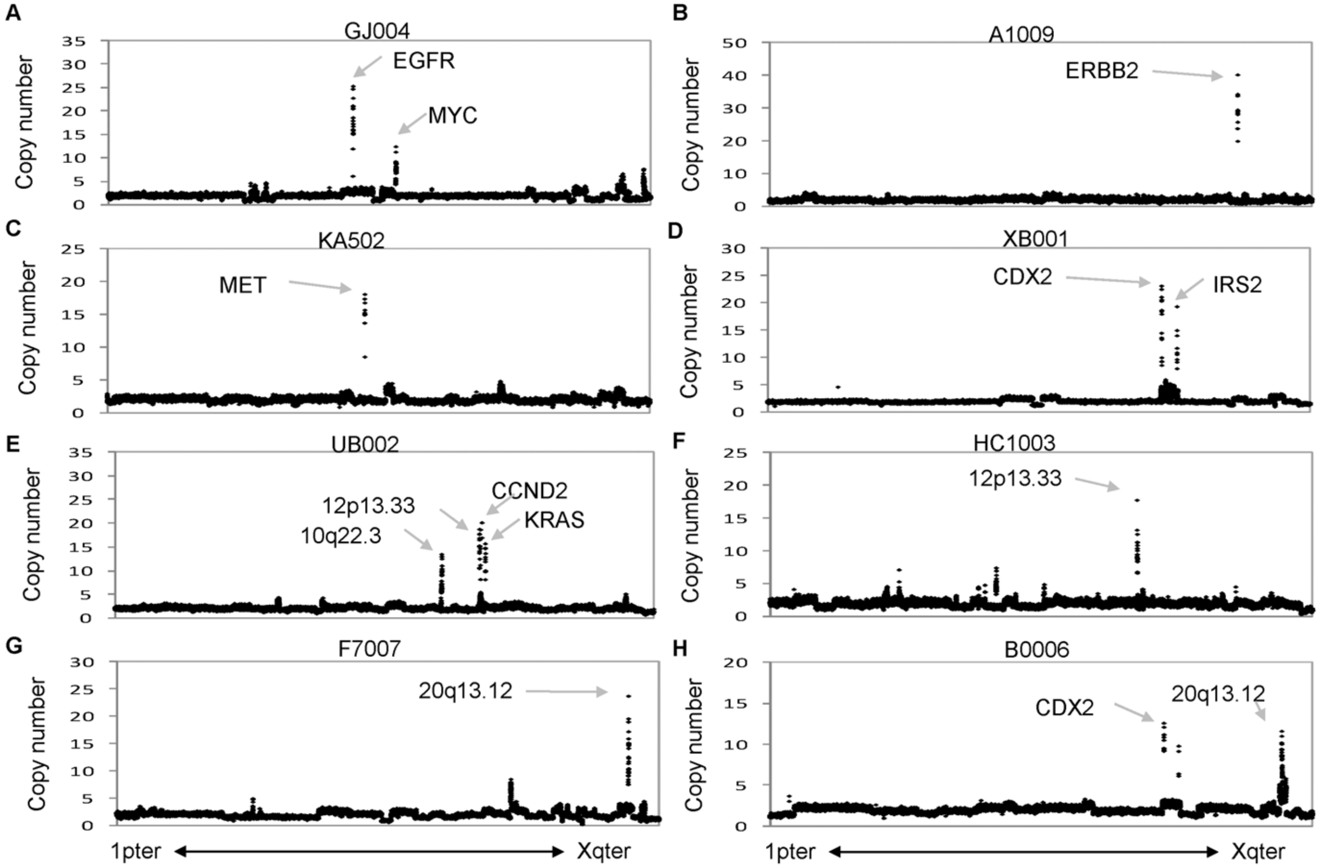


Figure 2. Focal amplification of genomic loci in selected stage II/III colon cancer samples. (**A–H**) Copy number plots for the entire genome arranged in chromosomal order from the short arm of chromosome 1 (1pter) to the long arm of chromosome X (Xqter) for 8 independent tumor samples. Amplicons of particular interest are highlighted with arrows, along with established oncogenes. Details regarding all amplicons and GISTIC peaks are in Table S3.
doi:10.1371/journal.pone.0042001.g002

To further search for patterns of affected pathway alterations, we mapped the list of genes amplified in CRC (**Table S4**) onto canonical molecular signaling pathways and cellular processes. **Table 1** shows top canonical pathways possibly affected by the amplified genes. Cell cycle is one of the most enriched pathways affected by somatic CNA involving genes such as CCND1, MYC, TFDP1 and YWHAZ. KEGG "Pathways in Cancer" underlies the broad spectrum effect of somatic CNAs in targeting multiple key pathways in cancer simultaneously. More specifically, we also identified individual cancer-related pathways that are significantly over-represented among cis-acting genes driven by somatic CNAs, including ERBB signaling pathway and MAPK kinase signaling pathway. Taken together, these results suggest that these somatic CNAs encode novel oncogenic driver genes and potential therapeutic targets in colon cancer.

CNA Clustering and Non-random CNA Correlations in CRC

We performed unsupervised hierarchical clustering of the global CNA data and identified three major clusters. Though we didn't find significant associations to age, gender, stage or KRAS mutation status, we observed that BRAF wild type tumors were significantly enriched in the largest cluster and BRAF mutants in one of the smaller clusters ($p<0.01$). Previously we [26] developed a BRAF-mutant gene expression signature from the PETACC-3 cohort and studied its prognostic implications. Among 213 MSS patients with mRNA expression data available, the signature identified 37 "BRAFm-like" samples (including 8 BRAF mutants) as well as 176 "non-BRAFm-like" samples. We re-ran clustering analysis on those 213 samples (**Figure 4A**), and found very significant enrichment of "non-BRAFm-like" samples ($p<0.01$) in the largest cluster (cluster 2) and "BRAFm-like" samples in cluster 1 ($P<0.01$, **Table 2**). Compared to cluster 2, cluster 1 shows much lower frequencies of amplification/deletion events, especially on chr13 q, 14 q, 18 q and 20 q (**Figure 4B**). A closer look reveals that cluster 1 is completely depleted from CNAs at chr20 while 95% of cluster 2 samples had chr20 amplified. These results corroborate with the observation of relative lower expression of chr20 genes in BRAFm-like with respect to the rest of the BRAFwt samples [26].

We previously reported that in cell lines CNAs at unlinked loci were frequently correlated to each other and that such correlations were likely the result of selection [31]. To assess whether a similar phenomenon was evident in clinical stage II/III MSS colon cancer, we conducted pair-wise correlations of copy number for all genes (~22 k) across the genome. As expected, adjacent (linked) genes were highly correlated (**Figure 5A**, close to diagonal). At a higher level some chromosome arms became unlinked (e.g. chr1p vs. 1 q, 10p vs. 10 q) or anti-correlated (e.g. chr8 p vs. 8 q). In addition, there were numerous correlations between unlinked loci (**Figure 5A**, off-diagonal), suggesting co-selection of these genomic regions. For example, chromosome 8 p losses were correlated to

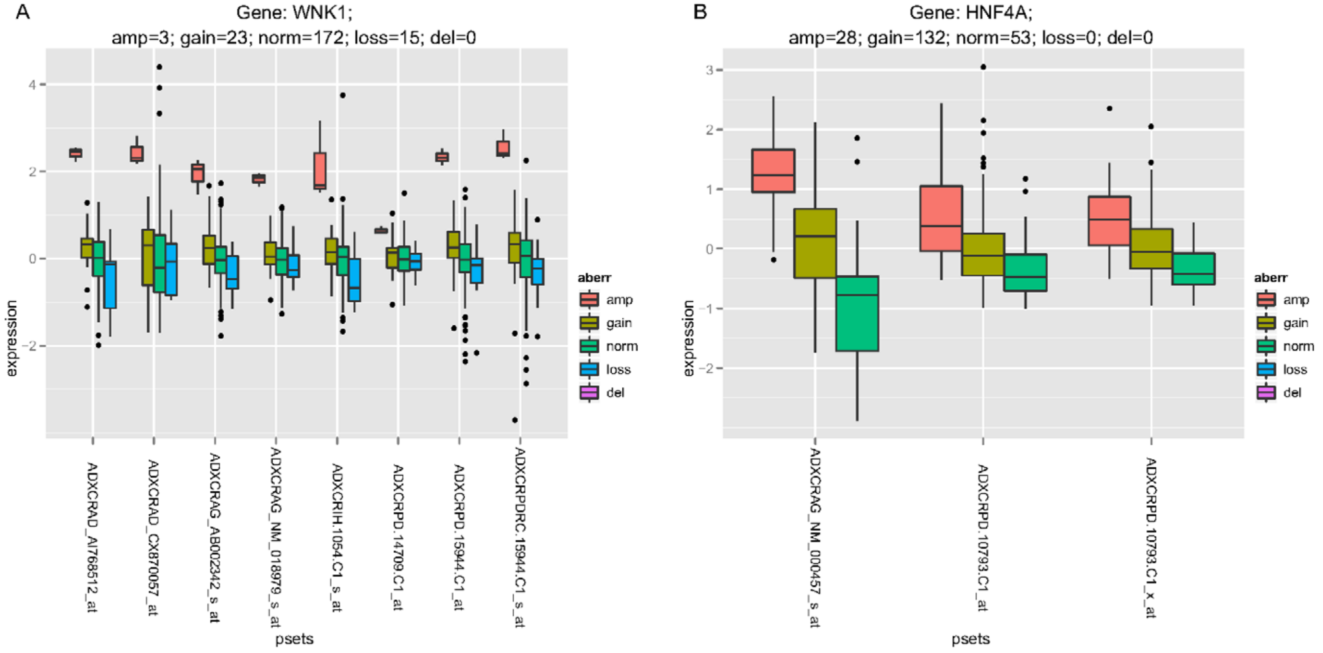


Figure 3. Boxplots for WNK1 (A) and HNF4A's (B) mRNA expression grouped by CNA status. Tumor samples were categorized by their CNA status (deletion, loss, normal, gain, amplification) for the indicated gene. The panels show the expression level by category for each probeset from the ALMAC platform (see Materials and Methods) representing the specific gene. The values were centered for each probeset; categories are plotted if there was at least one sample in it.
doi:10.1371/journal.pone.0042001.g003

losses of chromosomes 17 p and 18, along with gain of chromosome 20 q. Chromosome 13 gains were correlated to chromosome 14 losses. The distribution of gene-gene associations was significantly different than a randomization of the CNV data (**Figure 5B**). Similar to what was found in other cancer settings [31,34] there was a scale-free structure where a few genes were highly correlated to many other genes, while most genes correlated to only a few genes. This suggests that a small number of DNA loci act as hubs in a highly nonrandom hierarchical structure.

Relationship of CNA to Stage and Outcome

To identify individual CNAs that associate with tumor stage, we compared CNA frequencies between stage II (n = 30) and stage III MSS samples (n = 239). While both groups had similar patterns of CNA, a deletion on chromosome 3p14.2 had significantly ($p<0.01$) higher frequency in stage III tumors (24.3%) compared to stage II tumors (3.3%). This locus encodes FHIT, a candidate tumor suppressor and apoptotic regulator in colorectal cancer [35], and the higher frequency of deletion in stage III tumors suggests that loss of FHIT function may contribute to the progression of colon cancer from a lower to higher stage disease.

The large set of stage II/III MSS colon cancer samples with associated time-to-relapse, recurrence-free-survival (RFS) and overall survival (OS) afforded a unique opportunity to identify CNAs associated with outcome. Using Kaplan-Meier analysis, we first investigated whether the ch20q amplification revealed by sample clustering described previously lead to statistically significant differences in survival probability. A gained MCR on chromosome 20 q11.21-q13.33 (chr20:29,297,270–62,435,964 bp) was significantly associated with improved OS in stage III tumors ($p<0.01$). GISTIC identified one amplicon in this MCR on 20q13.33 (chr20:61,440,621–61,778,204 bp) which was

Table 1. Top canonical pathways possibly affected by the amplified genes.

Term	P-value	FDR*	Fold enrichment	% tumor amplified
KEGG_ADHERENS_JUNCTION	1.79E-04	4.50E-03	14.37	11.2%
KEGG_CELL_CYCLE	1.35E-03	2.01E-02	8.42	10.4%
KEGG_PATHWAYS_IN_CANCER	1.44E-03	2.32E-02	4.93	9.7%
KEGG_ERBB_SIGNALING_PATHWAY	3.16E-04	5.83E-03	12.39	8.2%
SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES	1.68E-03	2.42E-02	12.83	6.3%
BIOCARTA_TEL_PATHWAY	3.90E-05	2.13E-03	44.90	5.6%
KEGG_AXON_GUIDANCE	1.24E-02	7.48E-02	6.27	5.6%
KEGG_MAPK_SIGNALING_PATHWAY	1.77E-02	8.82E-02	4.04	5.6%

*FDR is based on was calculated based on 100 permutations where random sets of genes of same size were tested.
doi:10.1371/journal.pone.0042001.t001

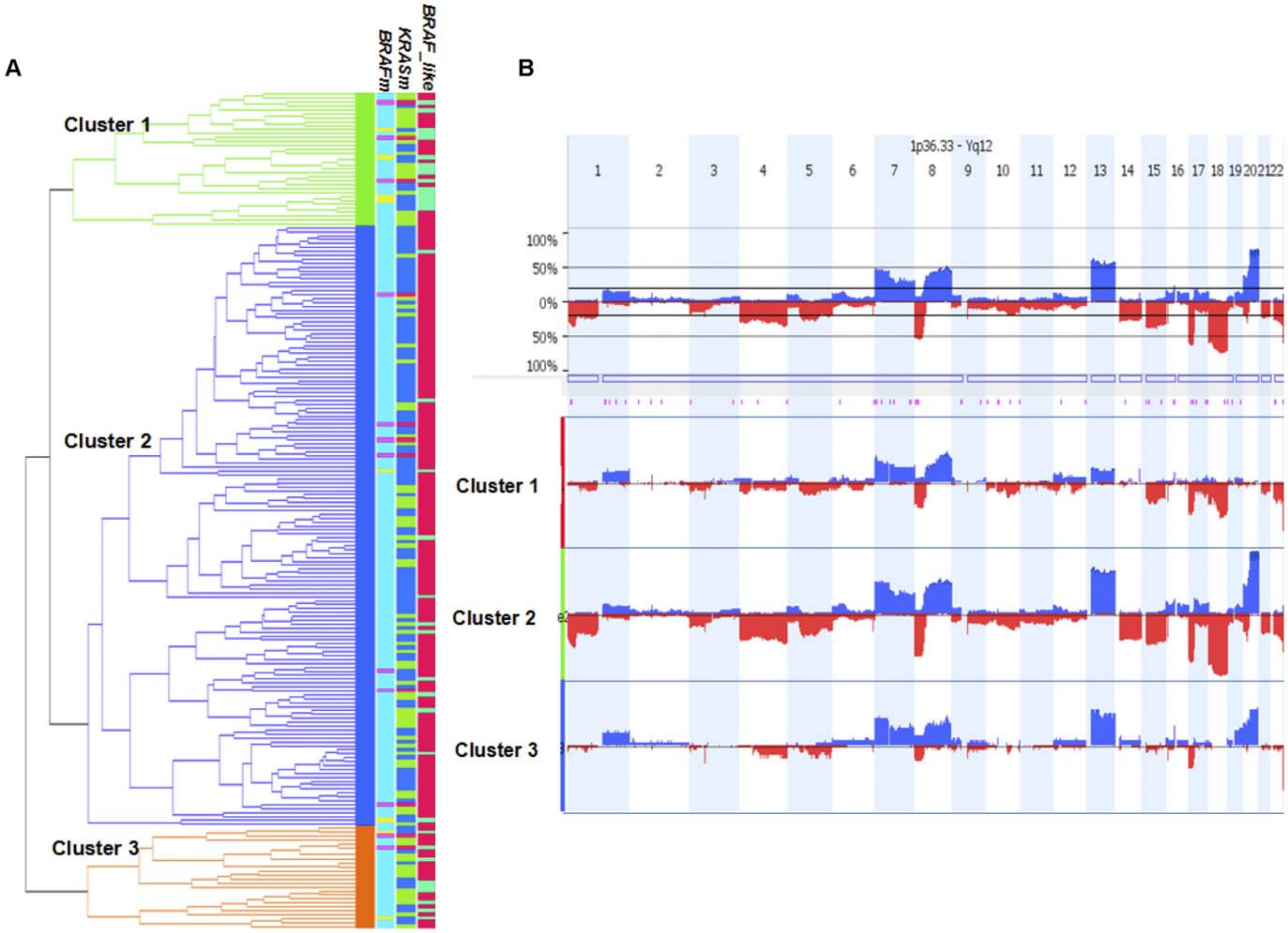


Figure 4. Unsupervised hierarchical clustering analysis based of genome-wide copy number data. (**A**) Three major clusters. The right-hand annotation indicates, in order, the BRAFm (in yellow, BRAF mutants; in blue, BRAF wild-types), KRASm (mutants in green), and BRAFm-like (in green, BRAFm-like; in red, non-BRAFm-like). Purple color indicates missing values. (**B**) Genome-wide frequency plot of copy number gain (above axis, blue) and copy number loss (below axis, red) across three major clusters.
doi:10.1371/journal.pone.0042001.g004

also significantly associated with improved OS. This region of approximately 300 kb contains one interesting genes such as EEF1A2 and PTK6. Anand et al. reported [36] EEF1A2's overexpression in about 30% of ovarian tumors and some established ovarian cancer cells. However, high EEF1A2 protein expression was associated with significantly increased 20-year survival probability in women with serous ovarian tumors [37], or in primary breast tumors, and this protective effect is thought to be due to EEF1A2's high expression in reducing the aggressiveness [38]. PTK6 was also reported [39] as positive associated to metastases-free survival in breast cancer; and shows strong cis CN/mRNA correlation in our analysis (T**able S4**). Here the CNA data suggest that amplification of the 20 q13.33 locus could be a significant prognostic marker of CRC cancer.

Table 2. Unsupervised hierarchical clustering indentified three major CNA clusters.

Cluster	samples	BRAFm-like	non-BRAF-like	BRAFm	BRAFwt	missing
1	34	16*	18	4*	27	3
2	153	12	141*	2	144*	7
3	26	9	17	2	22	2
Subtotal	213	37	176	8	193	12

*indicates significant over-representation in the category.
doi:10.1371/journal.pone.0042001.t002

Besides chr20q amplification, we applied Kaplan-Meier analysis to assess the relationship of all other MCRs and GISTIC peaks with RFS and OS. There were no significant associations between MCRs or GISTIC peaks versus OS or RFS for stage II tumors, possibly reflecting the limited number of samples in this group (n = 30). However, a deletion on chromosome 10 p (Chr10:0–10,743,764 bp) was significantly associated with poor RFS in stage III tumors alone ($p<0.01$) or stage II/III tumors combined ($p<0.01$), as well as poor OS in stage II/III tumors combined ($p<0.01$). Similarly, a deleted MCR on 19 p13.12 (chr19:14,425,490–15,580,441 bp) was significantly associated with OS ($p<0.01$) in stage II/III tumors combined (**Figure S2**).

Discussion

The main goals of this study were to develop a comprehensive overview of copy number aberrations (CNAs) and their associated

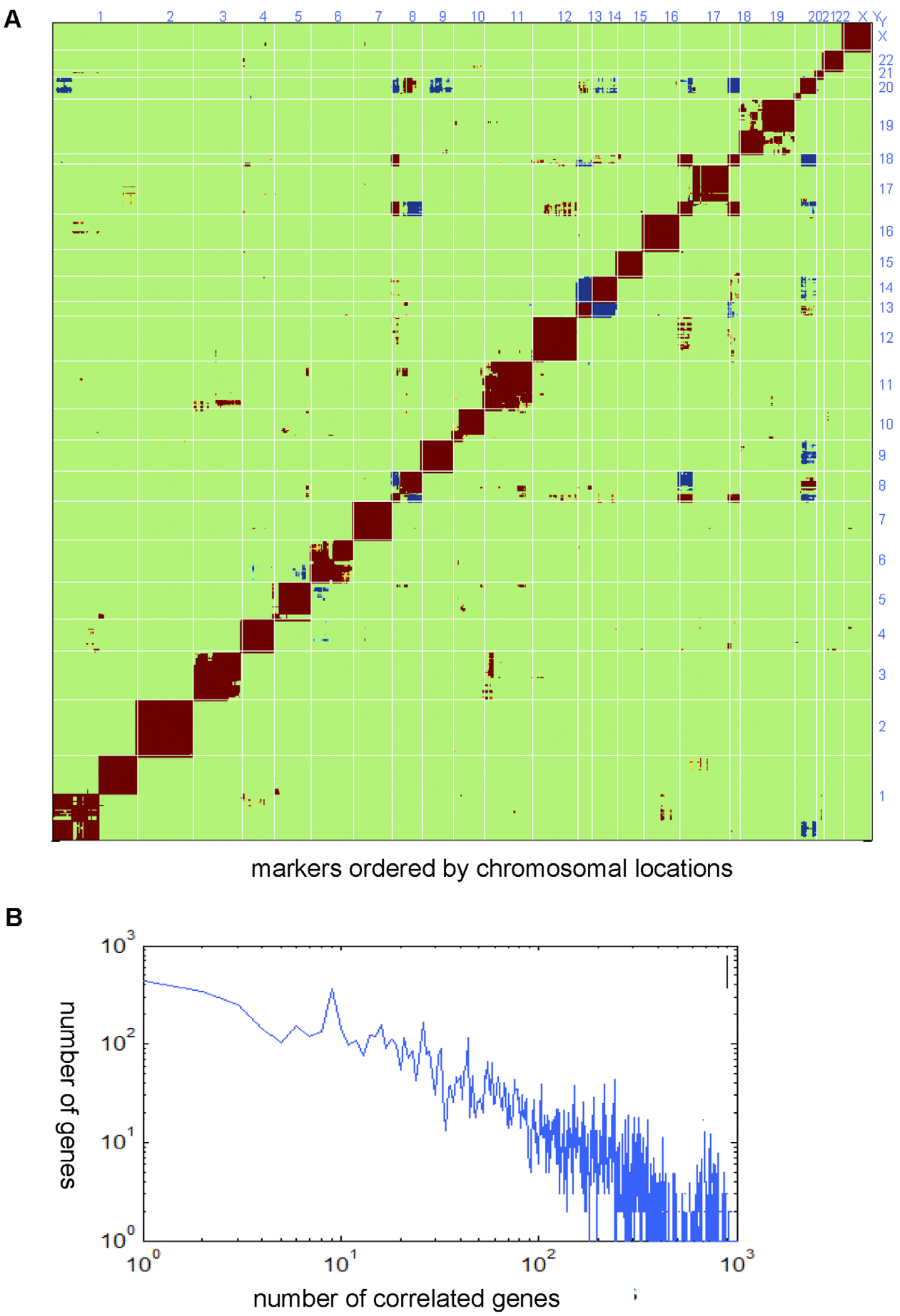

A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Y X 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
markers ordered by chromosomal locations
B
10^3
10^2
10^1
10^0
number of genes
10^0 10^1 10^2 10^3
number of correlated genes

Figure 5. Pair-wise DNA/DNA correlations reveal significant associations between unlinked loci. (**A**) Pair-wise correlations computed from gene copy number are ordered by chromosomal positions through the genome on the X and Y axes, with red indicating a positive correlation and blue indicating a negative correlation. The red diagonal represents the correlation of a gene with itself. The lower right and upper left portions of the graph represent mirror images of each other showing the copy number correlations of unlinked loci. (**B**) Log/log plots for significant gene/gene correlations (|R|≥0.3).
doi:10.1371/journal.pone.0042001.g005

genes in stage II/III colon cancer, to elucidate the underlying biology, and to associate CNAs with outcome. Regions of recurrent and focal CNA identified in these tumors highlight genomic regions most likely to encode oncogenes and tumor suppressors. Established oncogenes identified in this study that represent positive controls include MYC, CDX2, EGFR, MET, ERBB2, and CCND1.

The most prominent novel amplicons identified in this study include 12 p13.33 and multiple loci on 20 q (20 q11.21, 20 q13.12, 20 q13.31). The 12 p13.33 amplicon encodes the intriguing candidate WNK1, a member of the WNK family of serine/threonine kinases which affect MAPK signaling and a variety of cancer hallmarks including cell cycle progression, evasion of apoptosis, invasion and metastasis, and metabolic adaptation [40]. The complex pattern of gains and amplification on chromosome 20 q suggest multiple oncogenic drivers on this chromosome arm, consistent with observations in breast tumors [41]and other cancer types. The 20 q13.12 amplicon, which was observed in multiple tumors (**Figure 2G, 2H**) and is the most significant GISTIC peak on 20 q, encodes 11 genes, none of which have been unequivocally described as oncogenic drivers in colon cancer. Nonetheless, the reported functions of some of these genes suggest that further investigation is warranted. For example, the transcription factor HNF4A controls epithelial cell polarity and promotes gut neoplasia in mice [42]. WISP2 (WNT1 Inducible Signaling Pathway protein 2/CCN5) regulates the activity of the transforming growth factor â (TGFâ) signaling pathway and expression of genes associated with the epithelial-to-mesenchymal transition [43]. The peak at 20 q13.31 encodes BMP7, a member of the TGFâ superfamily of proteins whose overexpression in colorectal cancer significantly correlates with markers of pathological aggressiveness such as liver metastasis and is an independent prognostic factor of overall survival [44]. Functional characterization of these and other candidate oncogenes in colon cancer cell culture, patient-derived xenografts, or genetically engineered mouse models will help elucidate potential functional implications. Pathway analysis presented previously provides not only a better understanding of the possible biological context of candidate CNA drivers but also help to infer other genes on the altered pathway for which therapeutic options may be available. On the other hand, survival analysis shows improved overall survival for the sample segment with chr20 q13.33 amplification. This association contrasts with findings of another group who reported amplification of 20 q13 is indicating worse overall survival in sporadic colorectal cancers [45]. The exact basis for this discrepancy with our findings for is not clear, although the analyses of Aust et al. were on a substantially smaller cohort (120 samples).

Our analyses of associations between CNA and outcome in this set of stage II/III colon cancers revealed three loci that were significantly associated with overall survival (OS) or recurrence free survival (RFS). Deletion of the distal tip of chromosome 10 p (10 p15.3-p14) was associated with poor OS and RFS, while an interstitial deletion of chromosome 19 p (19 p13.12) was associated with poor OS, and gain of 20 q was associated with significantly better OS in stage III tumors. While 10 p deletions, 19 p deletions, and 20 q gains have been previously reported in stage II/III colon cancers [16], none of these loci have been previously linked to outcome in these tumors. Conversely, we did not observe significant associations of outcome to previously reported CNAs such as deletion of 16 p13.2 in stage II/III colon cancer [46], or deletion of 5 q34 and gain of 13 q22.1 in stage II tumors [17]. One potential explanation for these apparent discrepancies may relate to the limited power of the respective studies. For stage III MSS tumors, our results represent analyses of markedly higher sample numbers (n = 239) compared to published work (for e.g. 31 stage III tumors in [46]). For stage II MSS tumors, our sample set is underpowered, representing 30 samples compared to 41 [46] and 39 [17] tumors in earlier studies. These results emphasize the need for comprehensive analyses of large collections of clinically annotated tumor samples such as the stage III MSS tumor set described in this work.

We also reported here a significant non-random correlation of unlinked DNA loci with a scale-free structure in stage II/III colon cancer. These highly connected structures suggest a cycle of random changes in copy number followed by selection of a subset of changes that confer a selective advantage to tumor initiation and progression. While this is a long standing idea in cancer, correlation between unlinked loci suggests that highly ordered structures can emerge, potentially focused around biological functions of importance to the tumor. Future analyses could assess the effect of unlinked copy number correlations on gene expression, including enrichment of pathways and networks, and determining if the mRNA controlled by a pair of correlated loci overlap, where an independent effect of each loci was observable. This would identify pathways that were selectively altered during tumorigenesis and which therefore may represent new targetable functions.

Supporting Information

Figure S1 Boxplots for EGFR, ERBB2 and MYC's mRNA expression grouped by CNA status.
(PPT)

Figure S2 Kaplan-Meier curves demonstrate CNAs showing significant association with overall survival.
(PPT)

Table S1 Characteristics of patients Included in the study.
(XLS)

Table S2 Minimal common regions identified in 269 MSS stage-II/III colon cancer samples.
(XLS)

Table S3 GISTIC peaks identified in 269 MSS stage-II/III colon cancer samples.
(XLS)

Table S4 Affected genes in selected GISTIC amplicons.
(XLS)

Author Contributions

Conceived and designed the experiments: TX ST MD JGH. Analyzed the data: TX GA EM KW AFDN VP EB JGH. Contributed reagents/

materials/analysis tools: ST FTB ADR PY SB SLW. Wrote the paper: TX GA JRL ST MD VP MM PAR JGH.

References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, et al. (2008) Cancer Statistics, 2008. CA: A Cancer Journal for Clinicians 58: 71–96.
2. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. Cell 61: 759–767.
3. Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. Nat Med 10: 789–799.
4. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108–1113.
5. Speleman F, Kumps C, Buysse K, Poppe B, Menten B, et al. (2008) Copy number alterations and copy number variation in cancer: close encounters of the bad kind. Cytogenet Genome Res 123: 176–182.
6. Sayagues JM, Fontanillo C, Abad Mdel M, Gonzalez-Gonzalez M, Sarasquete ME, et al. (2010) Mapping of genetic abnormalities of primary tumours from metastatic CRC by high-resolution SNP arrays. PLoS ONE 5: e13752.
7. Ashktorab H, Schaffer AA, Daremipouran M, Smoot DT, Lee E, et al. (2010) Distinct genetic alterations in colorectal cancer. PLoS ONE 5: e8879.
8. Kurashina K, Yamashita Y, Ueno T, Koinuma K, Ohashi J, et al. (2008) Chromosome copy number analysis in screening for prognosis-related genomic regions in colorectal carcinoma. Cancer Sci 99: 1835–1840.
9. Platzer P, Upender MB, Wilson K, Willis J, Lutterbaugh J, et al. (2002) Silence of chromosomal amplifications in colon cancer. Cancer Res 62: 1134–1138.
10. Ried T, Knutzen R, Steinbeck R, Blegen H, Schrock E, et al. (1996) Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. Genes Chromosomes Cancer 15: 234–245.
11. Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, et al. (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer Res 64: 4817–4825.
12. Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, et al. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. Carcinogenesis 25: 1345–1357.
13. Camps J, Nguyen QT, Padilla-Nash HM, Knutsen T, McNeil NE, et al. (2009) Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. Genes Chromosomes Cancer 48: 1002–1017.
14. Ji H (2006) Molecular Inversion Probe Analysis of Gene Copy Alterations Reveals Distinct Categories of Colorectal Carcinoma. Cancer Research 66: 7910–7919.
15. Bartos JD, Gaile DP, McQuaid DE, Conroy JM, Darbary H, et al. (2007) aCGH local copy number aberrations associated with overall copy number genomic instability in colorectal cancer: coordinate involvement of the regions including BCR and ABL. Mutat Res 615: 1–11.
16. Reid JF, Gariboldi M, Sokolova V, Capobianco P, Lampis A, et al. (2009) Integrative approach for prioritizing cancer genes in sporadic colon cancer. Genes Chromosomes Cancer 48: 953–962.
17. Brosens RP, Haan JC, Carvalho B, Rustenburg F, Grabsch H, et al. (2010) Candidate driver genes in focal chromosomal aberrations of stage II colon cancer. J Pathol 221: 411–424.
18. Martin ES, Tonon G, Sinha R, Xiao Y, Feng B, et al. (2007) Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. Cancer Res 67: 10736–10743.
19. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, et al. (2009) Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. Proceedings of the National Academy of Sciences 106: 7131–7136.
20. Postma C, Koopman M, Buffart TE, Eijk PP, Carvalho B, et al. (2009) DNA copy number profiles of primary tumors as predictors of response to chemotherapy in advanced colorectal cancer. Annals of Oncology 20: 1048–1056.
21. Diep CB, Kleivi K, Ribeiro FR, Teixeira MR, Lindgjaerde OC, et al. (2006) The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. Genes Chromosomes Cancer 45: 31–41.
22. Venkatachalam R, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A, Kuiper RP (2008) Predisposition to colorectal cancer: exploiting copy number variation to identify novel predisposing genes and mechanisms. Cytogenet Genome Res 123: 188–194.
23. Nakao M, Kawauchi S, Furuya T, Uchiyama T, Adachi J, et al. (2009) Identification of DNA copy number aberrations associated with metastases of colorectal cancer using array CGH profiles. Cancer Genet Cytogenet 188: 70–76.
24. Poulogiannis G, Ichimura K, Hamoudi RA, Luo F, Leung SY, et al. (2010) Prognostic relevance of DNA copy number changes in colorectal cancer. J Pathol 220: 338–347.
25. Van Cutsem E, Labianca R, Bodoky G, Barone C, Aranda E, et al. (2009) Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. J Clin Oncol 27: 3117–3125.
26. Popovici V, Budinska E, Tejpar S, Weinrich S, Estrella H, et al. (2012) Identification of a Poor-Prognosis BRAF-Mutant-Like Population of Patients With Colon Cancer. J Clin Oncol.
27. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557–572.
28. Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, et al. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Res 16: 1149–1158.
29. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, et al. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 12: R41.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, et al. (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27: 1739–1740.
31. Xie T, Zhang C, Zhang B, Molony C, Oudes A, et al. (2010) A survey of cancer cell lines reveals highly structured and hierarchical relationships within and between DNA and mRNA that may be the result of selection. OMICS 14: 91–97.
32. Camps J, Armengol G, del Rey J, Lozano JJ, Vauhkonen H, et al. (2006) Genome-wide differences between microsatellite stable and unstable colorectal tumors. Carcinogenesis 27: 419–428.
33. Bouteille N, Driouch K, Hage PE, Sin S, Formstecher E, et al. (2009) Inhibition of the Wnt/beta-catenin pathway by the WWOX tumor suppressor protein. Oncogene 28: 2569–2580.
34. Lamb JR, Zhang C, Xie T, Wang K, Zhang B, et al. (2011) Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS ONE 6: e20090.
35. Wierzbicki PM, Adrych K, Kartanowicz D, Dobrowolski S, Stanislawowski M, et al. (2009) Fragile histidine triad (FHIT) gene is overexpressed in colorectal cancer. J Physiol Pharmacol 60 Suppl 4: 63–70.
36. Anand N, Murthy S, Amann G, Wernick M, Porter LA, et al. (2002) Protein elongation factor EEF1A2 is a putative oncogene in ovarian cancer. Nat Genet 31: 301–305.
37. Pinke DE, Kalloger SE, Francetic T, Huntsman DG, Lee JM (2008) The prognostic significance of elongation factor eEF1A2 in ovarian cancer. Gynecol Oncol 108: 561–568.
38. Kulkarni G, Turbin DA, Amiri A, Jeganathan S, Andrade-Navarro MA, et al. (2007) Expression of protein elongation factor eEF1A2 predicts favorable outcome in breast cancer. Breast Cancer Res Treat 102: 31–41.
39. Aubele M, Auer G, Walch AK, Munro A, Atkinson MJ, et al. (2007) PTK (protein tyrosine kinase)-6 and HER2 and 4, but not HER1 and 3 predict long-term survival in breast carcinomas. Br J Cancer 96: 801–807.
40. Moniz S, Jordan P (2010) Emerging roles for WNK kinases in cancer. Cell Mol Life Sci 67: 1265–1276.
41. Hodgson JG, Chin K, Collins C, Gray JW (2003) Genome amplification of chromosome 20 in breast cancer. Breast Cancer Res Treat 78: 337–345.
42. Darsigny M, Babeu JP, Seidman EG, Gendron FP, Levy E, et al. (2010) Hepatocyte nuclear factor-4alpha promotes gut neoplasia in mice and protects against the production of reactive oxygen species. Cancer Res 70: 9423–9433.
43. Sabbah M, Prunier C, Ferrand N, Megalophonos V, Lambein K, et al. (2011) CCN5, a novel transcriptional repressor of the transforming growth factor beta signaling pathway. Mol Cell Biol 31: 1459–1469.
44. Motoyama K, Tanaka F, Kosaka Y, Mimori K, Uetake H, et al. (2008) Clinical significance of BMP7 in human colorectal cancer. Ann Surg Oncol 15: 1530–1537.
45. Aust DE, Muders M, Kohler A, Schmidt M, Diebold J, et al. (2004) Prognostic relevance of 20 q13 gains in sporadic colorectal cancers: a FISH analysis. Scand J Gastroenterol 39: 766–772.
46. Andersen CL, Lamy P, Thorsen K, Kjeldsen E, Wikman F, et al. (2010) Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. Int J Cancer.

[*6*] Missiaglia E, Jacobs B, D'Ario G, Di Narzo AF, Soneson C, Budinska E, Popovici V, Vecchione L, Gerster S, Yan P, Roth AD, Klingbiel D, Bosman FT, Delorenzi M, Tejpar S. Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. Ann Oncol. 2014 Oct;25(10):1995-2001. doi: 10.1093/annonc/mdu275. Epub 2014 Jul 23. Erratum in: Ann Oncol. 2015 Feb;26(2):445. doi: 10.1093/annonc/mdu548. PMID: 25057166.

the Gynecological Cancer Intergroup (GCIG). Int J Gynecol Cancer 2011; 21: 419–423.

20. Markman M, Blessing J, Rubin SC et al. Phase II trial of weekly paclitaxel (80 mg/m2) in platinum and paclitaxel-resistant ovarian and primary peritoneal cancers: a Gynecologic Oncology Group study. Gynecol Oncol 2006; 101: 436–440.
21. Lortholary A, Largillier R, Weber B et al. Weekly paclitaxel as a single agent or in combination with carboplatin or weekly topotecan in patients with resistant ovarian cancer: the CARTAXHY randomized phase II trial from Groupe d'Investigateurs Nationaux pour l'Etude des Cancers Ovariens (GINECO). Ann Oncol 2011; 23: 346–352.
22. Pujade-Lauraine E, Hilpert F, Weber B et al. Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: the AURELIA open-label randomized phase III trial. J Clin Oncol 2014; 32: 1302–1308.
23. Baselga J, Cervantes A, Martinelli E et al. Phase I safety, pharmacokinetics, and inhibition of SRC activity study of saracatinib in patients with solid tumors. Clin Cancer Res 2010; 16: 4876–4883.
24. Hannon RA, Finkelman RD, Clack G et al. Effects of Src kinase inhibition by saracatinib (AZD0530) on bone turnover in advanced malignancy in a Phase I study. Bone 2012; 50: 885–892.
25. Linch M, Stavridi F, Hook J et al. Experience in a UK cancer centre of weekly paclitaxel in the treatment of relapsed ovarian and primary peritoneal cancer. Gynecol Oncol 2008; 109: 27–32.

Annals of Oncology 25: 1995–2001, 2014
doi:10.1093/annonc/mdu275
Published online 23 July 2014

Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features

E. Missiaglia[1], B. Jacobs[2], G. D'Ario[1], A. F. Di Narzo[1], C. Soneson[1], E. Budinska[3], V. Popovici[3], L. Vecchione[2], S. Gerster[1], P. Yan[4], A. D. Roth[5], D. Klingbiel[1,6], F. T. Bosman[4], M. Delorenzi[1,7,8] & S. Tejpar[2]*

[1]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [2]Molecular Digestive Oncology Unit, University Hospital Leuven, Leuven, Belgium; [3]Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic; [4]Department of Pathology, Lausanne University, Lausanne; [5]Oncosurgery Unit, Geneva University Hospital, Geneva; [6]The Swiss Group for Clinical Cancer Research (SAKK) Coordinating Center, Bern; [7]Ludwig Center for Cancer Research; [8]Oncology Department, University of Lausanne, Lausanne, Switzerland

Received 17 May 2014; revised 14 July 2014; accepted 15 July 2014

Background: Differences exist between the proximal and distal colon in terms of developmental origin, exposure to patterning genes, environmental mutagens, and gut flora. Little is known on how these differences may affect mechanisms of tumorigenesis, side-specific therapy response or prognosis. We explored systematic differences in pathway activation and their clinical implications.

Materials and methods: Detailed clinicopathological data for 3045 colon carcinoma patients enrolled in the PETACC3 adjuvant chemotherapy trial were available for analysis. A subset of 1404 samples had molecular data, including gene expression and DNA copy number profiles for 589 and 199 samples, respectively. In addition, 413 colon adenocarcinoma from TCGA collection were also analyzed. Tumor side-effect on anti-epidermal growth factor receptor (EGFR) therapy was assessed in a cohort of 325 metastatic patients. Outcome variables considered were relapse-free survival and survival after relapse (SAR).

Results: Proximal carcinomas were more often mucinous, microsatellite instable (MSI)-high, mutated in key tumorigenic pathways, expressed a B-Raf proto-oncogene, serine/threonine kinase (BRAF)-like and a serrated pathway signature, regardless of histological type. Distal carcinomas were more often chromosome instable and EGFR or human epidermal growth factor receptor 2 (HER2) amplified, and more frequently overexpressed epiregulin. While risk of relapse was not different per side, SAR was much poorer for proximal than for distal stage III carcinomas in a multivariable model including BRAF mutation status [$N = 285$; HR 1.95, 95% CI (1.6–2.4), $P < 0.001$]. Only patients with metastases from a distal carcinoma responded to anti-EGFR therapy, in line with the predictions of our pathway enrichment analysis.

Conclusions: Colorectal carcinoma side is associated with differences in key molecular features, some immediately druggable, with important prognostic effects which are maintained in metastatic lesions. Although within side significant

**Correspondence to:* Prof. Sabine Tejpar, Molecular Digestive Oncology Unit, University Hospital Leuven, 3000 Leuven, Belgium. Tel: +32-16-34-42-25; Fax: +32-16-34-44-19; Email: sabine.tejpar@uzleuven.be

molecular heterogeneity remains, our findings justify stratification of patients by side for retrospective and prospective analyses of drug efficacy and prognosis.

Key words: colon cancer, expression profiling, mutations, oncogenic pathways, survival

introduction

Current understanding of molecular mechanisms involved in colorectal cancer (CRC) supports three main molecular pathways. The almost classical chromosomal instability (CIN) pathway is based on the seminal publication of Vogelstein and contains most of the kirsten rat sarcoma viral oncogene homolog (KRAS) mutated CRCs. The mismatch repair deficient or microsatellite instable (MSI) pathway was discovered through elucidation of the gene mutations responsible for Lynch syndrome and is characterized by a hypermutating state and frequent B-Raf proto-oncogene, serine/threonine kinase (BRAF) V600E mutation. The CpG island methylator phenotype (CIMP) pathway goes along with the occurrence of serrated precursor lesions and is also strongly related to the MSI pathway, notably through frequent methylation of the mutL homolog 1 promoter, which confers MSI-high status [1]. The pathways are sufficiently distinct to be conceptually valid, but they also significantly overlap. This makes the development of new molecular modalities of classification of CRC a complex task [2].

Different approaches toward molecular classification have been undertaken, based on gene expression profiles and the TCGA whole-genome sequencing effort. We and others have proposed gene expression-based molecular subgroups [3–6] that share (groups of) molecular characteristics while maintaining significant intragroup heterogeneity. Typical examples are the segregation of clinically significant subgroups such as those BRAF-mutated or expressing a BRAF-mutated gene expression signature [7] and MSI or expressing an MSI-like signature [8]. Signatures and subgroups identified by them intend to define patient categories for which treatment needs and/or response to treatment may differ.

The systematic attempt toward subclassification is epitomized in the TNM staging approach and stage grouping as its derivative. Anatomic characteristics related to tumor spread still dictate to a large extent, even in this era of molecular scrutiny, how a patient will be treated. Strikingly, tumor side in terms of proximal or distal colon has gained in prominence in recent years. Initially, this was recognized mostly through the strong preference for the proximal colon for cancers associated with the Lynch syndrome. This paved the way toward the recognition that proximal carcinomas are more often MSI, BRAF-mutated and express the CIMP phenotype [9, 10]. This might be related to differences in biology between the proximal and distal colon, with potentially significant impact on tumorigenesis in these respective sides. However, little is known about the mechanisms responsible for such tumor heterogeneity. One distinctive feature is represented by their embryonic derivation, which is the midgut and the hindgut for the proximal and distal colon, respectively. The pathways involved in the development of these segments have been extensively explored and should be taken into consideration when the biology of their derived cancers is considered. Additionally, the differences in luminal content and bacterial flora between the left and right colon may influence oncogenesis [11]. Therefore, tumor location is a major source of biological heterogeneity, potentially with prognostic and predictive implications in view of the fact that the mortality rate is higher in proximal than in distal colon cancer (CC) [12–15].

We hypothesized that the carcinogenic pathway is different between proximal and distal colon tumors, and that this would be reflected in size-associated differences in the molecular characteristics of the tumors. This might have profound prognostic and therapeutic implications. We tested this by comparing clinicopathological and molecular characteristics of carcinomas in the proximal versus distal colon in two large CC cohorts.

materials and methods

patients

Clinicopathological data were available for a cohort of 3045 CC patients enrolled in the PETACC3 adjuvant chemotherapy trial. A subset of those patients had molecular data ($N = 1404$), including BRAF, KRAS, and PIK3CA mutation status, MSI status, and 18q arm loss of heterozygosity (LOS). Parallel gene expression ($N = 589$) and DNA copy number profiles ($N = 199$) were also available [16, 17]. Clinicopathological ($N = 413$) and molecular information (somatic mutations $N = 199$, RNAseq $N = 325$) for additional CC patients were obtained from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) [18].

Gene expression profiles of 84 normal colon samples were derived from four datasets (TCGA CC, GSE14333, GSE8671, and GSE41258).

To assess tumor side-effect on response to anti-epidermal growth factor receptor (EGFR) therapy, we studied a cohort of 435 chemorefractory metastatic CRC patients [19].

Tumors located in the splenic flexure, descending colon, and sigmoid colon were defined as proximal, while cecum, ascending, and hepatic flexure were classified as distal. Intraperitoneal rectum and distal rectum were excluded from the analysis. Transverse CCs (for the lack of clarity as to the exact location) were included exclusively when assessing feature distribution along the bowel. Further information is given in supplementary Materials and Methods, available at *Annals of Oncology* online.

statistical analysis

Gene expression and copy number data analyses were processed as described elsewhere [3, 16]. Biological interpretation was carried out using tools and signatures described in supplementary Materials and Methods, available at *Annals of Oncology* online. We applied a Bayesian model selection approach to test if variables could be explained better by a flat, dichotomous, or a continuum model of variation along the bowel.

We assessed differences in the distribution of categorical variables with Fisher's test or Pearson's χ^2 test, as indicated. We used the Cox proportional hazards model to assess the association of tumor side with time-to-event end points and Kaplan–Meier method for figures.

results

The frequency distribution of the clinicopathological features along the bowel was analyzed using the PETACC3 and TCGA cohorts. Proximal carcinomas were associated with higher age, node-negative stage, high grade, and mucinous differentiation (supplementary Table S1 and S2, available at *Annals of Oncology* online). Furthermore, proximal carcinomas disseminated more often to the abdominal viscera and lymph nodes, whereas distal carcinomas had a higher frequency of liver and chest metastases. Concerning the distributions of the variables along the bowel, including, for this, the transverse colon (supplementary Table S3 and S4, available at *Annals of Oncology* online) most of them favored a biphasic model, with the exception of MSI in the PETACC3 dataset which showed a gradual distribution. Based on these findings, we explored the molecular bases of such differences starting from the colon normal mucosa.

Gene expression profiles of 84 normal samples (34 proximal and 50 distal) collected from four public datasets were analyzed to assess the effect on gene expression in normal mucosa based on their location. In a meta-analytical approach including colon side as a predictor, we identified 351 genes differentially expressed—157 overexpressed in the proximal and 194 in the distal colon (supplementary Table S5, available at *Annals of Oncology* online). Notably, the expression of some HOX genes involved in colon development (HOXC6, HOXB6, and HOXB13) as well as of the EGFR ligand epiregulin (EREG) was different according to side. Gene set enrichment analysis using DAVID evidenced that genes overexpressed in the proximal colon were associated with an inflammatory response and drug metabolism (notably of cytochrome P450 superfamily—supplementary Table S6, available at *Annals of Oncology* online).

Difference in gene expression between proximal and distal tumors was explored in 589 CC samples (211 proximal and 378 distal) from the PETACC3 dataset, using a linear model controlling for potential confounders such as BRAF and KRAS mutation status and MSI. After correction for multiple testing, 576 genes were found differentially expressed (158 genes up-regulated in proximal and 418 in distal carcinomas—supplementary Table S7, available at *Annals of Oncology* online), showing mainly a biphasic midgut/hindgut pattern, as for the clinicopathological features. Overall, gene expression fold-changes between the two sides were small in magnitude.

Only 20 genes (including two HOX genes—HOXC6 and HOXB13) were found to be in common with the 351 genes found differentially expressed in the normal colon. Notably, within the group of BRAF-mutated carcinomas (which are mostly proximal), no differences were found between proximal and distal carcinomas (data not shown).

To elucidate if tumor side influences the type of pathways exploited by tumor cells to promote and sustain CC tumorigenesis, we selected a set of gene signatures representing the main biological processes involved in CC (details in supplementary Table S8, available at *Annals of Oncology* online). The level of those signatures was compared between sides in 589 CC from the PETACC3 dataset and 325 from the TCGA dataset and results combined meta-analytically. Figure 1 summarizes the strength and direction of the association between

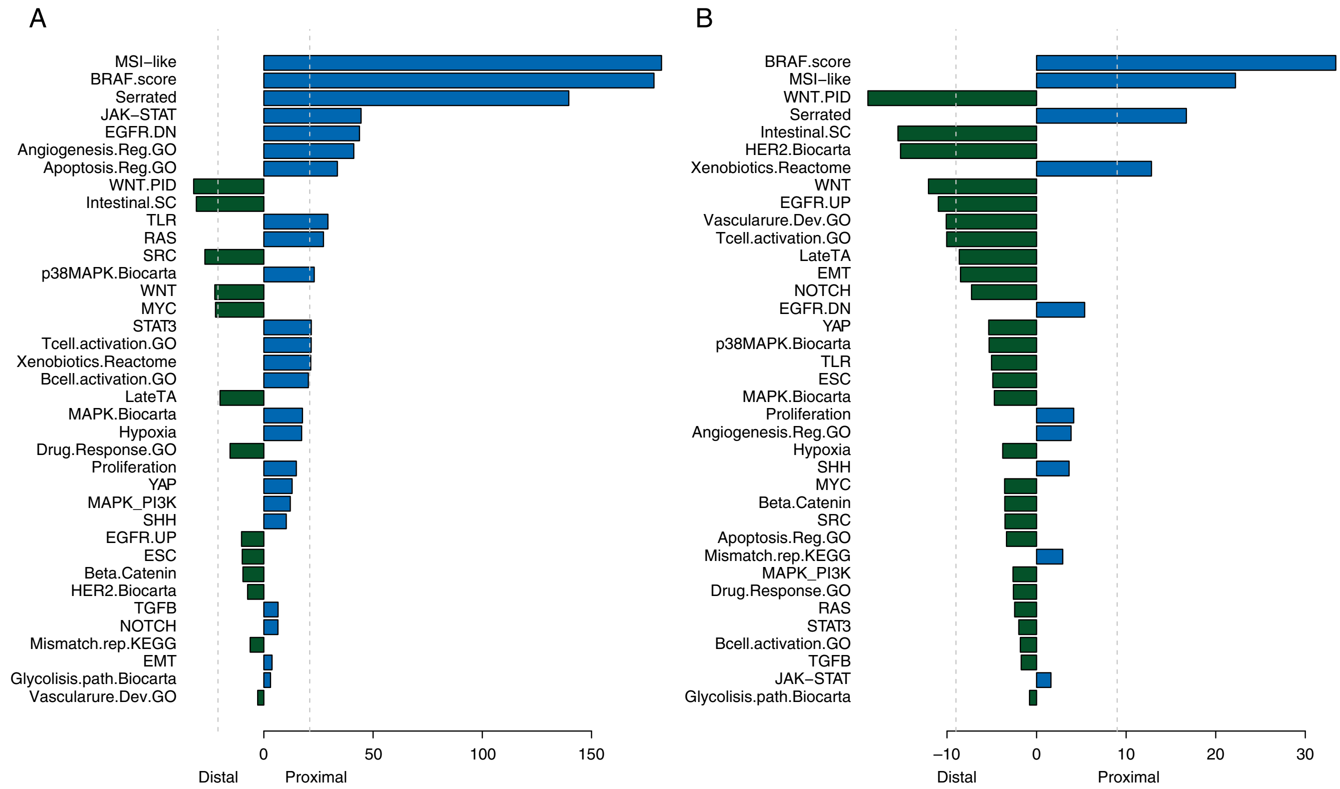


Figure 1. Barplot showing signed statistic of the association between gene signatures and tumor side observed in 589 CRC from the PETACC3 and 325 from the TCGA datasets. (A) The analysis was carried out considering all the patients, (B) or focusing on MSS, BRAF, and KRAS wild-type tumors. Association of the gene signatures with tumor location was assessed separately within each dataset using a linear model. Results were combined using Fisher's method. Blue bars represent levels of significance after adjustment for multiple testing [$P < 0.05$ after Bonferroni correction for all patients (A) and false discovery rate (FDR) <0.25 after Benjamini–Hochberg procedure when considering MSS, BRAF, and KRAS wild-type tumors (B)].

the signatures and tumor side considering all samples or focusing exclusively on microsatellite stable (MSS), BRAF, and KRAS wild-type patients. BRAF-like, MSI-like, and serrated adenoma signatures showed the strongest bias between sides, suggesting that these are the most prevalent signatures distinguishing proximal from distal tumors. Notably, this difference is also observed in the set of MSS, KRAS, and BRAF wild-type tumors (supplementary Figure S1, available at *Annals of Oncology* online). In the whole patient cohort, we also found a significant positive association between proximal tumors and T-cell activation, JAK-STAT, angiogenesis, apoptosis, RAS, and mitogen-activated protein kinase (MAPK) activation. In contrast, distal carcinomas were associated with WNT, MYC, and SCR activation as well as the presence of intestinal stem cells. Notably, distal MSS and BRAF and KRAS wild-type carcinomas were also associated with human epidermal growth factor receptor 2 (HER2) and EGFR activation signaling, which parallels the observation that EREG (EGF ligand) was among the most overexpressed genes in distal carcinomas.

Copy number variation (CNV) analysis was carried out on a subset of 199 patients (127 distal and 72 proximal) from the PETACC3 study. Distal carcinomas showed a significantly higher proportion of CIN+ patients (57%) than proximal carcinomas (40%) (χ^2 test, $P = 0.029$), as well as a higher number of amplification/deletions (supplementary Figure S2, available at *Annals of Oncology* online). Regions on chromosomes 10, 11, 14, 18, and 20 were altered with different frequency (supplementary Table S9 and Figure S3 and S4, available at *Annals of Oncology* online). Notably, gain of 20q and loss of 18q were found significantly more often in distal carcinomas (supplementary Figure S2 and Table S1, available at *Annals of Oncology* online), which corroborate overexpressed in distal tumors of a significant proportion (20%) of genes located on 20q (Fisher's test, $P < 0.0001$).

Chromosomal regions hosting receptor tyrosine kinases were more often amplified in distal (60/127, 47%) than in proximal (23/72; 32%) carcinomas, including the ErbB family members HER2 and EGFR (16/127 versus 1/72, Fisher's test $P < 0.001$; supplementary Figure S5, available at *Annals of Oncology* online).

Mutation frequency was analyzed in 199 tumors (78 distal and 121 proximal) from the TCGA CC collection. As previously described [18], mutations were more frequent in MSI-high than in MSS carcinomas (data not shown). However, in proximal MSS carcinomas, the number of deleterious mutations was higher than in distal MSS carcinomas (supplementary Figure S6, available at *Annals of Oncology* online), even after removing all hypermutant tumors (non-silent mutation rate >450). A similar trend was also observed when considering only oncogenes, indicating that the higher mutation rate was potentially an important feature of proximal tumors beyond the MSI/hypermutated status.

This was confirmed by the observation that important signaling pathways such as MAPK, ErbB, TGF-beta, and insulin signaling pathways were found more frequently mutated in proximal than in distal carcinomas (supplementary Table S10, available at *Annals of Oncology* online). As supportive evidence, we found a similar mutation bias in the PETACC3 dataset for oncogenes, such as BRAF, KRAS, and PIK3Ca (supplementary Table S1, available at *Annals of Oncology* online).

We explored the association of tumor side with relapse-free survival (RFS) and survival after relapse (SAR) in the PETACC3 cohort. Surprisingly, stage II proximal carcinomas relapsed significantly less frequently than those in the distal colon (supplementary Figure S7, available at *Annals of Oncology* online). However, this appeared to be entirely due to the MSI population (mostly proximal), as this was no longer found when only MSS carcinomas were considered. For stage III patients, no effect of side was found on RFS (supplementary Figure S8, available at *Annals of Oncology* online). Multivariable analysis confirmed that side is not an independent prognostic factor for RFS (supplementary Table S11, available at *Annals of Oncology* online).

In contrast, when stage III patients with a proximal carcinoma became metastatic, they had a significantly worse survival than those with a metastatic distal carcinoma [HR 1.97, 95% CI (1.6–2.3), $P < 0.001$; supplementary Figure S7, available at *Annals of Oncology* online]. Multivariable analysis showed that this effect was independent of MSI and KRAS or BRAF mutation status [HR 1.7, 95% CI (1.3–2.4), $P < 0.001$; supplementary Table S11, available at *Annals of Oncology* online]. The BRAF signature score, which is higher in proximal carcinomas and itself highly prognostic for SAR [7], outcompeted side in a multivariable model (data not shown), although in the non-BRAF mutant-like subset side was still a significant factor (supplementary Figure S8, available at *Annals of Oncology* online).

In the smaller stage II proximal carcinoma cohort, we also observed a trend toward poorer outcome. This was confirmed in an independent untreated population (supplementary Figure S8, available at *Annals of Oncology* online).

In view of our finding that, in distal tumors, the frequency of amplification of ErbB family members is higher and the activation of EGFR signaling stronger, we explored if EGFR inhibitor efficacy is affected by tumor side. To this end, we studied 435 metastatic chemorefractory patients (126 or 29% proximal and 309 or 71% distal), of whom 207 were KRAS and BRAF wild-type (WT2) and had been treated with cetuximab combined with chemotherapy [19].

Overall, in univariable models, patients with a distal carcinoma showed better progression-free survival [PFS; 21 weeks (95% CI 19–24 weeks)] than those with a proximal carcinoma [13 weeks (95% CI 11–17 weeks); $P < 0.001$; supplementary Figure S9, available at *Annals of Oncology* online]. This was largely due to patients with a WT2 carcinoma, of whom the median PFS was 18 weeks in case of a proximal carcinoma (95% CI 11–31 weeks) but 30 weeks in case of a distal carcinoma (95% CI 26–34 weeks, $P = 0.02$). In contrast, KRAS or BRAF-mutated carcinomas did not show any difference in outcome according to side (data not shown).

discussion

It is now clear that CRC is a molecularly heterogeneous disease [3–6], and that this heterogeneity should be used to stratify patients for optimal response to current and novel therapeutic strategies. We confirm the emerging notion that a significant part of this heterogeneity is captured by the anatomic location of the tumor. However, we were not able to confirm that those changes occur gradually along the bowel, as previously hypothesized [10].

We found differences in gene expression between the proximal and distal normal colon, which mostly overlapped with those found by LaPointe et al. [20], but which did not emerge as significant in the differences between proximal and distal carcinomas.

We confirm that proximal tumors are more often MSI and hypermutated, which is at least in part due to their deficient DNA mismatch repair status. However, in both the PETACC3 and TCGA series, even non-hypermutant proximal MSS carcinomas harbor more potentially deleterious mutations, including mutations of *KRAS*, *BRAF*, and *PIK3Ca*. We observed a higher frequency of BRAF-mutated, BRAF score, and serrated signature expressing proximal carcinomas, as was also found in mouse models recapitulating human BRAF^{V600E} mutated serrated lesions with an MSI phenotype [21]. Proximal carcinomas, often characteristically mucinous, densely infiltrated with tumor-infiltrating lymphocytes, and with activated MAPK signaling, might develop from precursor lesions driven by pathways which are associated with side-specific cellular characteristics, such as tolerance to DNA repair defects and to oncogenic stress. In addition, environmental factors like bacterial toxins or mutagenic CYP450 metabolites, which increase the mutation rate, may contribute to the specific characteristics of these cancers [11].

In contrast, distal carcinomas characteristically harbor numerous large chromosomal alterations (notably gain of 20q and loss of 18q), for which the responsible mechanisms are not fully understood. Loss of 18q [22] as well as activation of EGF signaling, which induce the expression of *AURKA* [23], might be implicated. We found HER1 and HER2, directly druggable targets, amplified in 12% of distal carcinomas (9% of which wild-type for *KRAS* and *BRAF*) and gene expression evidence of activation of the EGFR pathway largely restricted to the distal colon. The observation that, in the adenomatous polyposis coli mouse model, the disruption of the pan-ErbB-negative regulator LRIG1 predominantly induces distal neoplasms [24] supports the hypothesis of an important contribution of EGF signaling to distal colon carcinogenesis.

These differences in mutation rate and genomic instability between the two colon sides are striking and need to be better understood both in terms of their bearing on prognosis as well as response to DNA repair targeting chemotherapies. Multivariable analyses, containing all major known risk factors including *BRAF* and *KRAS* mutations, showed that side is an independent prognostic factor for SAR. Furthermore, *BRAF* mutant or BRAF-

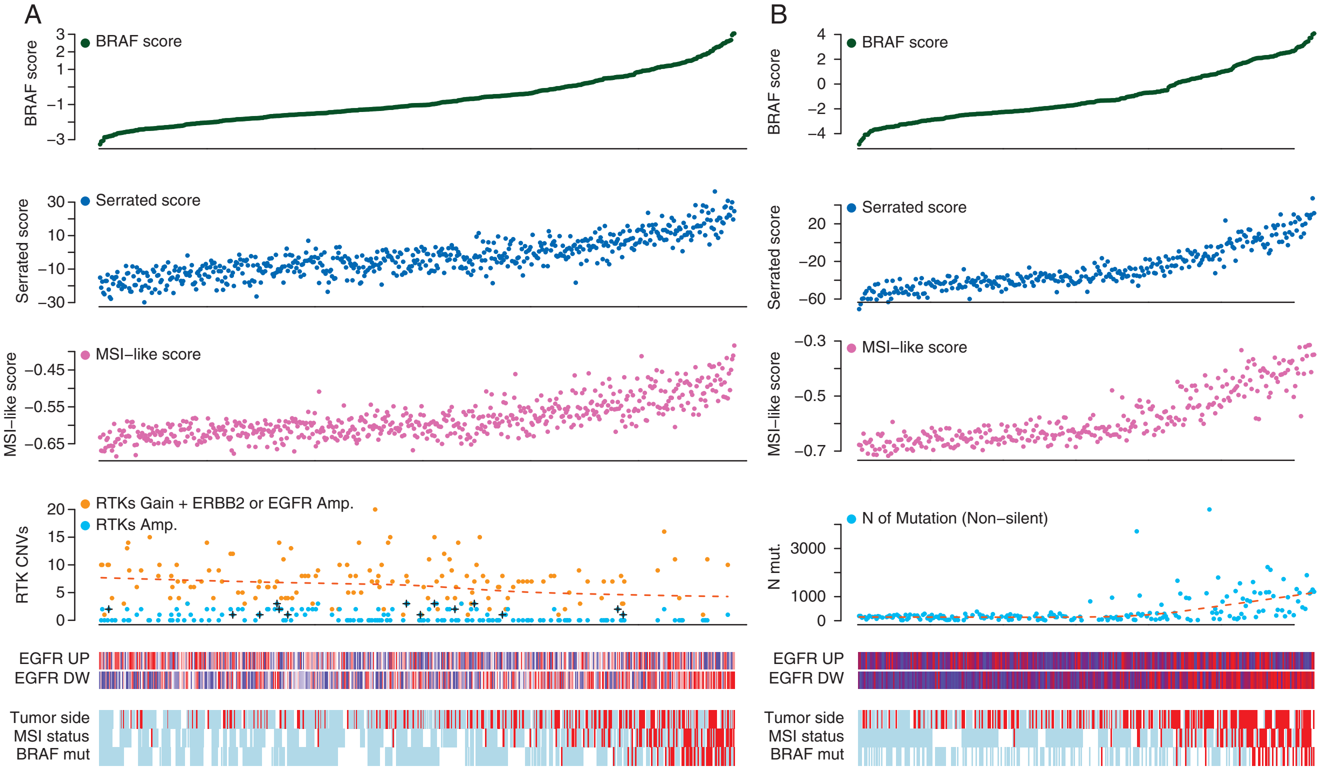


Figure 2. Distribution of key variables ranked by BRAF score level in PETACC3 ($N = 589$) (A) and TCGA colon adenocarcinoma collection ($N = 314$) (B). From top to bottom panel, the plots depict: (1) ranking of the BRAF score, (2) serrated adenoma score, and (3) MSI-like score. These scores showed a strong association in both datasets. (4) Copy number variations of chromosomal locus of 43 receptor tyrosine kinases (RTKs) (PETACC3 dataset). The number of RTKs per patients showing gain (one copy) and amplification (more than two copies) is plotted. Patients showing amplification of HER1 and/or HER2 are also marked. A smooth curve was fitted by Loess (local polynomial regression fitting) using smoother span of 60%. In TCGA, we plotted the number of non-silent mutations observed per patient. A smooth curve was also fitted. (5) The median expression of the genes included in the EGFR signature set from Kobayashi et al. (UP and DW) (supplementary Materials and Methods, available at *Annals of Oncology* online) is reported as heatmap. (6) Distribution of tumors by side (red—proximal and blue—distal), MSI (red—MSI-H and blue—MSS), and BRAF status (red—BRAF mutant and blue—BRAF wild-type).

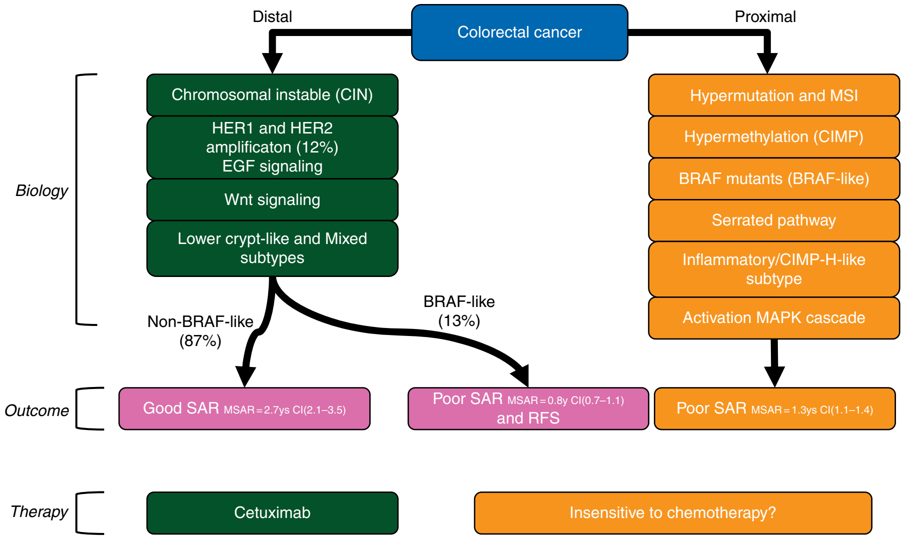

Distal
Colorectal cancer
Proximal
Biology
Chromosomal instable (CIN)
HER1 and HER2 amplificaton (12%) EGF signaling
Wnt signaling
Lower crypt-like and Mixed subtypes
Hypermutation and MSI
Hypermethylation (CIMP)
BRAF mutants (BRAF-like)
Serrated pathway
Inflammatory/CIMP-H-like subtype
Activation MAPK cascade
Non-BRAF-like (87%)
BRAF-like (13%)
Outcome
Good SAR MSAR = 2.7ys CI(2.1–3.5)
Poor SAR MSAR = 0.8y CI(0.7–1.1) and RFS
Poor SAR MSAR = 1.3ys CI(1.1–1.4)
Therapy
Cetuximab
Insensitive to chemotherapy?

Figure 3. Summary of the main biological and clinical findings. SAR, survival after relapse; MSAR, median survival after relapse; y, year.

like distal carcinomas have poorer SAR and RFS [7]. We hypothesize that metastases of proximal colon carcinomas have an increased mutation rate and higher cellular plasticity, potentially exacerbated by the effects of chemotherapy, with as a potential consequence a deleterious effect of (neo)adjuvant therapy. The combination of hypermethylation and a hypermutant state may induce, in metastases of proximal carcinomas, resistance to the current, mostly 5-fluorouracil-based, chemotherapeutic regimens. Our current working hypothesis is that proximal carcinomas have a poor prognosis under current best care, which should be confirmed by reanalysis by tumor side of all major CRC trials. Efficacious treatment of proximal carcinomas might require completely different drug regimens.

Our observations of an active EGFR signaling in distal carcinomas also suggest that those tumors benefit significantly more from anti-EGFR agents than proximal carcinomas, which was supported by our results obtained from a single-arm study. This finding also emerged recently from the NCIC-CTG-CO.17 reanalysis of cetuximab monotherapy versus best supportive care and emphasizes that benefit is restricted to proximal carcinomas [25].

In summary, the molecular and clinical characteristics of proximal and distal colon carcinomas are significantly different (as is summarized in Figures 2 and 3) and show to go beyond the simple MSI–MSS grouping. It remains to be seen if the findings hold also in advanced diseases, under-represented in our study. Tumor location is yet another simplistic subdivision of CRC, but it does go along with significant and characteristic molecular heterogeneity based on differences in biology, which is potentially highly relevant for therapeutic decision-making.

acknowledgements

We thank all the clinicians who enrolled patients and participated in the PETACC-3 trial (see Appendix at: http://jco.ascopubs.org/content/27/19/3117.long), in particular the coordinators David Cunningham, Roberto Labianca and Eric Van Cutsem.

funding

This work was supported by the grants SNF 320030_135421 of the Swiss National Science Foundation and KFS 0269708-2010 of the Krebsliga Schweiz to AR and MD, by the EU Seventh Framework Programme, grant agreement 259015 COLTHERES to MD and ST, and by grants from the Belgian National Cancer Plan (Nationaal Kankerplan) and from the Fondation Roi Baudoin to ST, who is also supported as a Senior Clinical Investigator of the Fund for Scientific Research Flanders (FWO-Vlaanderen). BJ is supported as a Postdoctoral fellow of the Fund for Scientific Research Flanders (FWO-Vlaanderen). We thank the Fondation Medic for financial support.

disclosure

AR: advisor Pfizer. ST: speaker fee Merck Serono, advisor Merck Serono and Sanofi, and past research grant Pfizer. The others authors have declared no conflicts of interest.

references

1. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 2007; 50(1): 113–130.
2. Snover DC. Update on the serrated pathway to colorectal carcinoma. Hum Pathol 2011; 42(1): 1–10.
3. Budinska E, Popovici V, Tejpar S et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 2013; 231(1): 63–76.
4. De Sousa EMF, Wang X, Jansen M et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med 2013; 19(5): 614–618.
5. Marisa L, de Reynies A, Duval A et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013; 10(5): e1001453.
6. Sadanandam A, Lyssiotis CA, Homicsko K et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 2013; 19(5): 619–625.
7. Popovici V, Budinska E, Tejpar S et al. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. J Clin Oncol 2012; 30(12): 1288–1295.

8. Tian S, Roepman P, Popovici V et al. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. J Pathol 2012; 228(4): 586–595.
9. Sanz-Pamplona R, Cordero D, Berenguer A et al. Gene expression differences between colon and rectum tumors. Clin Cancer Res 2011; 17(23): 7303–7312.
10. Yamauchi M, Morikawa T, Kuchiba A et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. Gut 2012; 61(6): 847–854.
11. Iacopetta B. Are there two sides to colorectal cancer? Int J Cancer 2002; 101(5): 403–408.
12. Benedix F, Kube R, Meyer F et al. Comparison of 17,641 patients with right- and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival. Dis Colon Rectum 2010; 53(1): 57–64.
13. Meguid RA, Slidell MB, Wolfgang CL et al. Is there a difference in survival between right- versus left-sided colon cancers? Ann Surg Oncol 2008; 15(9): 2388–2394.
14. Weiss JM, Pfau PR, O'Connor ES et al. Mortality by stage for right- versus left-sided colon cancer: analysis of surveillance, epidemiology, and end results—Medicare data. J Clin Oncol 2011; 29(33): 4401–4409.
15. Sinicrope FA, Mahoney MR, Smyrk TC et al. Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy. J Clin Oncol 2013; 31(29): 3664–3672.
16. Xie T, G DA, Lamb JR et al. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. PLoS ONE 2012; 7(7): e42001.
17. Bosman FT, Yan P, Tejpar S et al. Tissue biomarker development in a multicentre trial context: a feasibility study on the PETACC3 stage II and III colon cancer adjuvant treatment trial. Clin Cancer Res 2009; 15(17): 5528–5533.
18. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487(7407): 330–337.
19. De Roock W, Claes B, Bernasconi D et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncol 2010; 11(8): 753–762.
20. LaPointe LC, Dunne R, Brown GS et al. Map of differential transcript expression in the normal human large intestine. Physiol Genomics 2008; 33(1): 50–64.
21. Rad R, Cadinanos J, Rad L et al. A genetic progression model of Braf(V600E)-induced intestinal tumorigenesis reveals targets for therapeutic intervention. Cancer Cell 2013; 24(1): 15–29.
22. Burrell RA, McClelland SE, Endesfelder D et al. Replication stress links structural and numerical cancer chromosomal instability. Nature 2013; 494(7438): 492–496.
23. Lai CH, Tseng JT, Lee YC et al. Translational up-regulation of Aurora-A in EGFR-overexpressed cancer. J Cell Mol Med 2010; 14(6B): 1520–1531.
24. Powell AE, Wang Y, Li Y et al. The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. Cell 2012; 149 (1): 146–158.
25. Brule SY, Jonker DJ, Karapetis CS et al. Location of colon cancer (right-sided [RC] versus left-sided [LC]) as a predictor of benefit from cetuximab (CET): NCIC CTG CO.17. J Clin Oncol 2013: abstract 31, (suppl; abstr 3528).

Annals of Oncology 25: 2001–2007, 2014
doi:10.1093/annonc/mdu361
Published online 31 July 2014

Fourfold increased detection of Lynch syndrome by raising age limit for tumour genetic testing from 50 to 70 years is cost-effective

A. S. Sie[1], A. R. Mensenkamp[1], E. M. M. Adang[2], M. J. L. Ligtenberg[1,3] & N. Hoogerbrugge[1*]

Departments of [1]Human Genetics; [2]Health Evidence; [3]Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

Received 20 March 2014; revised 11 June 2014; accepted 21 July 2014

Background: Recognising colorectal cancer (CRC) patients with Lynch syndrome (LS) can increase life expectancy of these patients and their close relatives. To improve identification of this under-diagnosed disease, experts suggested raising the age limit for CRC tumour genetic testing from 50 to 70 years. The present study evaluates the efficacy and cost-effectiveness of this strategy.

Methods: Probabilistic efficacy and cost-effectiveness analyses were carried out comparing tumour genetic testing of CRC diagnosed at age 70 or below (experimental strategy) versus CRC diagnosed at age 50 or below (current practice). The proportions of LS patients identified and cost-effectiveness including cascade screening of relatives, were calculated by decision analytic models based on real-life data.

Results: Using the experimental strategy, four times more LS patients can be identified among CRC patients when compared with current practice. Both the costs to detect one LS patient (€9437/carrier versus €4837/carrier), and the number needed to test for detecting one LS patient (42 versus 19) doubled. When family cascade screening was

*Correspondence to: Prof. Nicoline Hoogerbrugge, Department of Human Genetics 836, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, The Netherlands. Tel: +31-24-366-62-05; E-mail: nicoline.hoogerbrugge@radboudumc.nl

Update

Annals of Oncology

Volume 26, Issue 2, February 2015, Page 445

DOI: https://doi.org/10.1093/annonc/mdu548

Annals of Oncology 26: 445, 2015
doi:10.1093/annonc/mdu547

Phase I/IIa study evaluating the safety, efficacy, pharmacokinetics, and pharmacodynamics of lucitanib in advanced solid tumors

J.-C. Soria[1], F. DeBraud[2], R. Bahleda[1], B. Adamo[3], F. Andre[1], R. Dienstmann[3,4], A. Delmonte[2], R. Cereda[5,6,7], J. Isaacson[5,6,7], J. Litten[5,6,7], A. Allen[5,6,7], F. Dubois[8], C. Saba[8], R. Robert[8], M. D'Incalci[9], M. Zucchetti[9], M. G. Camboni[5,6,7] & J. Tabernero[3]

[1]Department of Drug Development, Gustave-Roussy Cancer Campus, Villejuif, France; [2]European Institute of Oncology, Milan, Italy; [3]Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron University Hospital, Universitat Autònoma de Barcelona, Barcelona, Spain; [4]Sage Bionetworks, Fred Hutchinson Cancer Research Center, Seattle; [5]Clovis Oncology, Inc., San Francisco; [6]Clovis Oncology, Inc., Boulder, USA; [7]Clovis Oncology, Inc., Milan, Italy; [8]Institut de Recherche International Servier, Suresnes, France; [9]Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa, Milan, Italy

Ann Oncol 2014; 25: 2244–2251 (doi:10.1093/annonc/mdu390)

There was a spelling error in the author list in the original manuscript. The correct authors and affiliations are as above. The authors apologize for the errors.

Annals of Oncology 26: 445, 2015
doi:10.1093/annonc/mdu548

Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features

Ann Oncol 2014; 25: 1995–2001 (doi: 10.1093/annonc/mdu275)

There were some errors in the original manuscript. These have been corrected below. The authors apologize for the errors.

abstract-results

Proximal carcinomas were more often mucinous, microsatellite instable (MSI)-high, mutated in key tumorigenic pathways, expressed a B-Raf proto-oncogene, serine/threonine kinase (BRAF)-like and a serrated pathway signature, regardless of histological type. Distal carcinomas were more often chromosome instable and EGFR or human epidermal growth factor receptor 2 (HER2) amplified, and more frequently overexpressed epiregulin. While risk of relapse was not different per side, SAR was much poorer for proximal than for distal stage III carcinomas in a multivariable model including BRAF mutation status [N = 285; HR 1.95, 95% CI (1.6–2.4), $P < 0.001$]. Only patients with metastases from a distal carcinoma responded to anti-EGFR therapy, in line with the predictions of our pathway enrichment analysis.

materials and methods

patients

Clinicopathological data were available for a cohort of 3045 CC patients enrolled in the PETACC3 adjuvant chemotherapy trial. A subset of those patients had molecular data (N = 1404), including BRAF, KRAS, and PIK3CA mutation status, MSI status, and 18q arm loss of heterozygosity (LOS). Parallel gene expression (N = 589) and DNA copy number profiles (N = 199) were also available [16, 17]. Clinicopathological (N = 413) and molecular information (somatic mutations N = 199, RNAseq N = 325) for additional CC patients were obtained from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) [18].

Gene expression profiles of 84 normal colon samples were derived from four datasets (TCGA CC, GSE14333, GSE8671, and GSE41258). To assess tumor side-effect on response to anti-epidermal growth factor receptor (EGFR) therapy, we studied a cohort of 435 chemorefractory metastatic CRC patients [19].

Tumors located in the splenic flexure, descending colon, and sigmoid colon were defined as distal, while cecum, ascending, and hepatic flexure were classified as proximal. Intraperitoneal rectum and distal rectum were excluded from the analysis. Transverse CCs (for the lack of clarity as to the exact location) were included exclusively when assessing feature distribution along the bowel. Further information is given in supplementary Materials and Methods, available at *Annals of Oncology* online.

[*7*] Popovici V, **Budinska E,** Bosman FT, Roth AD, Delorenzi M, Tejpar S. Identification of a Poor-Prognosis BRAF-Mutant-Like Population of Patients With Colon Cancer. *J Clin Oncol.* 2012 Oct 1;30(28):128-43. doi:10.1200/JCO.2011.41.1607.

Identification of a Poor-Prognosis BRAF-Mutant-Like Population of Patients With Colon Cancer

Article *in* Journal of Clinical Oncology · March 2012

DOI: 10.1200/JCO.2011.39.5814 · Source: PubMed

Vlad Popovici
Masaryk University
155 PUBLICATIONS **4,305** CITATIONS
SEE PROFILE

Eva Budinská
Masaryk University
157 PUBLICATIONS **5,101** CITATIONS
SEE PROFILE

Sabine Tejpar
KU Leuven
525 PUBLICATIONS **42,514** CITATIONS
SEE PROFILE

John Graeme Hodgson
Ariad Pharmaceuticals
82 PUBLICATIONS **14,087** CITATIONS
SEE PROFILE

VOLUME 30 · NUMBER 12 · APRIL 20 2012

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Identification of a Poor-Prognosis *BRAF*-Mutant–Like Population of Patients With Colon Cancer

Vlad Popovici, Eva Budinska, Sabine Tejpar, Scott Weinrich, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Tao Xie, Fred T. Bosman, Arnaud D. Roth, and Mauro Delorenzi

See accompanying editorial on page 1255; listen to the podcast by Dr Meyerhardt at www.jco.org/podcasts

Vlad Popovici, Eva Budinska, and Mauro Delorenzi, Swiss Institute of Bioinformatics; Fred T. Bosman and Mauro Delorenzi, Lausanne University Medical Center, Lausanne; Arnaud D. Roth, Geneva University Hospital, Geneva; Arnaud D. Roth, The Swiss Group for Clinical Cancer Research, Bern, Switzerland; Sabine Tejpar and Eric Van Cutsem, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium; and Scott Weinrich, Heather Estrella, Graeme Hodgson, and Tao Xie, Pfizer, La Jolla, CA.

Submitted September 23, 2011; accepted December 20, 2011; published online ahead of print at www.jco.org on March 5, 2012.

Supported by the Swiss National Center for Competence in Research Molecular Biology (V.P. and M.D.) and Fondation Medic (E.B.); S.T. is a senior clinical investigator of the Fund for Scientific Research Flanders and has received research grants from the Belgian Federation Against Cancer and from the Belgian National Cancer Plan.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Vlad Popovici, PhD, Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Quartier Sorge, Genopode, CH-1015 Lausanne, Switzerland; e-mail: vlad.popovici@isb-sib.ch.

0732-183X/12/3012-1288/$20.00

DOI: 10.1200/JCO.2011.39.5814

ABSTRACT

Purpose

Our purpose was development and assessment of a *BRAF*-mutant gene expression signature for colon cancer (CC) and the study of its prognostic implications.

Materials and Methods

A set of 668 stage II and III CC samples from the PETACC-3 (Pan-European Trails in Alimentary Tract Cancers) clinical trial were used to assess differential gene expression between c.1799T>A (p.V600E) *BRAF* mutant and non-*BRAF*, non-*KRAS* mutant cancers (double wild type) and to construct a gene expression–based classifier for detecting *BRAF* mutant samples with high sensitivity. The classifier was validated in independent data sets, and survival rates were compared between classifier positive and negative tumors.

Results

A 64 gene-based classifier was developed with 96% sensitivity and 86% specificity for detecting *BRAF* mutant tumors in PETACC-3 and independent samples. A subpopulation of *BRAF* wild-type patients (30% of *KRAS* mutants, 13% of double wild type) showed a gene expression pattern and had poor overall survival and survival after relapse, similar to those observed in *BRAF*-mutant patients. Thus they form a distinct prognostic subgroup within their mutation class.

Conclusion

A characteristic pattern of gene expression is associated with and accurately predicts *BRAF* mutation status and, in addition, identifies a population of *BRAF* mutated-like *KRAS* mutants and double wild-type patients with similarly poor prognosis. This suggests a common biology between these tumors and provides a novel classification tool for cancers, adding prognostic and biologic information that is not captured by the mutation status alone. These results may guide therapeutic strategies for this patient segment and may help in population stratification for clinical trials.

J Clin Oncol 30:1288-1295. © 2012 by American Society of Clinical Oncology

INTRODUCTION

Activation of the *KRAS/BRAF/MEK/ERK* cascade is believed to occur frequently in colorectal (CRC) cancer on the basis of the observed 40% incidence of *KRAS* mutations and 10% to 15% incidence of *BRAF* mutations.[1-4] *KRAS* and *BRAF* mutations occur in a mutually exclusive pattern in CRC, which has long been interpreted as a sign of functional redundancy. However, these mutations occur in different histopathologic subtypes of CRC,[5,6] and we recently showed[7] that the prognosis of patients with *KRAS* and *BRAF* mutant metastatic CRC is quite different, with a clearly worse prognosis for *BRAF*-mutant disease. It has been suggested this could be due to higher levels of mitogen-activated protein kinase activation in *BRAF*-mutant (BRAFm) colon cancer.[8,9] Unlike the majority of *KRAS*-mutant (KRASm) CRCs, BRAFm metastatic CRCs do not respond to any current chemotherapy, and the outcome of patients with BRAFm CRC is similar to that of untreated patients.

Our main objective was to better understand the underlying biology of BRAFm CRCs as captured by gene expression. We developed a BRAFm gene signature that allowed an accurate identification of BRAFm samples, and which, when applied to *BRAF* wild-type samples, identified additional colon cancer (CC) samples that manifested a similar gene expression pattern. Although a substantial amount of work has been dedicated to the development of BRAFm gene expression signatures in melanoma,[10-12] to the best of our knowledge, there is no such published work in the CC context. Taking advantage of a large series of tumors with gene expression and mutation data from the PETACC-3 (Pan-European Trails in Alimentary Tract Cancers) clinical trial,[13] we studied the genes differentially expressed between c.1799T>A (p.V600E) BRAFm and double-wild-type (WT2) tumors, defined as non-*BRAF* mutant, non-*KRAS* mutant. We purposely excluded the KRASm tumors from this comparison because it was unclear whether KRASm carcinomas had overlapping biology with BRAFm. Next, we built a classifier able to recognize with high sensitivity BRAFm CCs in our own and external data sets.

When the *BRAF* classifier was applied to the whole population, it identified a *BRAF* wild-type subpopulation, with similar gene expression and prognostic characteristics. Approximately 62% of these BRAFm-like tumors were KRASm (30% of all KRASm were BRAFm-like), with the rest being WT2 (13% of all WT2). In our data, the BRAFm-like population represented 18% of CCs. This intriguing finding suggests a common biology between these tumors, not predicted by the mutation status. The results obtained show that our current classifications of tumors as *KRAS*- or *BRAF*-mutant or mitogen-activated protein kinase–active versus nonactive are inadequate to capture the whole underlying biology and clinical behavior.

MATERIALS AND METHODS

Tumor Samples and Data Preparation

Within the PETACC-3 clinical trial,[13] formalin-fixed paraffin-embedded tissue blocks were collected after cancer diagnosis and independently of future research plans, and DNA was extracted from 1,404 microdissected tissue sections. The analysis of *KRAS* exon 2 and *BRAF* exon 15 was performed by allele-specific real-time polymerase chain reaction.[7] The mutation status has been confirmed for all samples by a second analysis, using Sequenom.[14] RNA of sufficient quantity and quality was extracted from 895 samples, and gene expressions were measured on the ALMAC Colorectal Cancer DSA platform (Craigavon, Northern Ireland)—a customized Affymetrix chip with 61,528 probe sets mapping to 15,920 unique Entrez Gene IDs—in two phases (phase 1: n = 322, phase 2: n = 573). In total, 688 unique samples passed the final quality control (phase 1: n = 265 [82.3%], phase 2: n = 423 [73.8%]) and were used in subsequent analysis (Data Supplement). Of this series of CCs, 257 (37.4%) were *KRAS* mutated, whereas *BRAF* mutation was detected in 47 (6.8%) of the cases (Data Supplement).

The stage III subset included all samples for which profile data could be obtained and is thus representative of the clinical population of the trial. The stage II subset included all patients with relapse for whom profile data could be obtained and is thus also representative of this group, whereas from the nonrelapsing patients, a randomly selected population was profiled.

Three additional independent data sets[15-17] were used for validation of the signature, whereas a fourth data set,[18] with available survival information, was used for validating the prognostic value of the signature.

Statistical Analysis

PETACC-3 gene expression data were retrospectively analyzed to derive the *BRAF* gene signature discriminating between c.1799T>A (p.V600E) BRAFm and double-wild-type (WT2; *BRAF* and *KRAS* wild-type) tumors. Samples with missing mutation information (n = 39) were discarded from the gene signature development, but were included later in the survival analysis.

Gene expression data were normalized using RMA (Robust Microchip Average)[19] and summarized at the gene level by choosing the probe set with the highest standard deviation as a representative of each gene, in each data set individually.

Differentially expressed genes were obtained by fitting multivariate linear models (using LIMMA[20] package) to probe set–level data to fully exploit the potential of the platform. To account for known association between microsatellite instability-high (MSI-H), BRAFm, and right-sided tumors,[7] the linear model for the whole population included factors for *BRAF* mutation, MSI status, and tumor site (all binary variables). For the microsatellite stable (MSS) subpopulation, the model included only the *BRAF* mutation status and tumor site. The false discovery rate was controlled by Benjamini-Hochberg procedure[21] and required to be at most 1%, whereas the minimum absolute log-fold change was 0.585 (= log2 1.5). As the MSI-H subpopulation was small and consisted only of right-sided samples, the differentially expressed genes were derived by comparing BRAFm and WT2 only in the right colon, with a false discovery rate less than 25% and no constraint on the fold change.

For signature generation, an adapted version of the top scoring pairs algorithm[22] (multiple top scoring pairs [mTSP]; Data Supplement) was used, resulting in gene pairs deemed as the most informative in the process of classifier construction. The final classification model consisted of two groups of genes (G1 and G2), and the prediction was made comparing the averages of these groups: If, for a given sample, the average of G1 was smaller than the average of G2, then the sample was predicted to be BRAFm, otherwise WT2.

We also defined a *BRAF* score (BS) as the difference between the average expression of G2 genes and the average expression of G1 genes (from the mTSP model) and used it to analyze the stratification for different threshold values (a threshold of 0 leading to the original decision rule). An alternative threshold for the *BRAF* score was obtained as the value that maximized Matthews correlation coefficient[23] on the PETACC-3 data set.

The performance of the classifier was estimated by repeated (10 times) stratified five-fold cross-validation, following the MAQC-II guidelines,[24] and measured in terms of sensitivity, specificity, and error rate. The final *BRAF* classifier was built from all BRAFm and WT2 samples in the PETACC-3 data set and then applied to the full PETACC-3 data set (including KRASm) and independent validation sets for the analysis of stratification of the population (Data Supplement). Because the stage II subgroup of PETACC-3 is smaller and not fully representative, the analysis of the prognostic value of the signature is focused on stage III subgroup. However, results for both stages are given (Data Supplement).

The association between predicted class and survival outcomes was tested using Cox proportional hazard models (log-likelihood test) and log-rank test for dichotomous variables. Three survival outcomes have been considered: overall survival, relapse-free survival and survival after relapse. Fisher's exact test was used for testing differences in proportions in contingency tables.

RESULTS

BRAFm: Characteristic Genes and Classifier

In the PETACC-3 data set, we identified 314 differentially expressed probe sets between BRAFm and WT2 (see Materials and Methods for details), mapping to 223 unique EntrezGene IDs. Top 50 differentially expressed probe sets are given in Table 1, with the full table given in the Data Supplement. We also derived lists of differentially expressed genes for the MSI-H and MSS tumors separately (Data Supplement).

Using the technique of mTSP, a 32-gene pair BRAFm signature (Table 2) was obtained by training on the c.1799T>A (p.V600E) BRAFm and WT2 samples, considering all genes, whether or not they were previously identified to be differentially expressed. Its performance was estimated at a sensitivity of 95.8% and a specificity of 86.5% (Table 3). Fifty of the 64 genes of the signature were among the 223 differentially expressed genes (Data Supplement).

BRAFm-Like Tumors

To make the distinction between the true and classifier-predicted mutation status, we prefix the predictions by "pred-": pred-BRAFm denotes the samples predicted to be BRAFm, whereas pred-BRAFwt

Table 1. Top 50 Differentially Expressed Probe Sets Between c.1799T>A (p.V600E) BRAFm and WT2

Probe Set ID	Gene Symbol	Entrez GeneID	LFC	Official Full Name
ADXCRPD.7995.C1_x_at	AQP5	362	−2.91	Aquaporin 5
ADXCRIH.384.C1_s_at	REG4	83998	−2.80	Regenerating islet-derived family, member 4
ADXCRAG_BC014461_x_at	CDX2	1045	2.02	Caudal type homeobox 2
ADXCRAG_BC014461_at	CDX2	1045	1.97	Caudal type homeobox 2
ADXCRPD.10572.C1_at	HSF5	124535	1.70	Heat shock transcription factor family member 5
ADXCRAG_AK024491_s_at	SOX8	30812	−1.95	SRY (sex determining region Y)-box 8
ADXCRSS.Hs#S2988180_at	HSF5	124535	2.02	Heat shock transcription factor family member 5
ADXCRPD.7687.C1_at	TM4SF4	7104	−1.70	Transmembrane 4 L six family member 4
ADXCRAG_M14335_s_at	F5	2153	−1.18	Coagulation factor V (proaccelerin, labile factor)
ADXCRAG_AJ250717_s_at	CTSE	1510	−2.62	Cathepsin E
ADXCRAG_AJ132099_s_at	VNN1	8876	−0.93	Vanin 1
ADXCRAD_NM_025113_s_at	C13orf18	80183	1.77	Chromosome 13 open reading frame 18
ADXCRAG_NM_182510_s_at	LOC146336	146336	−1.33	Hypothetical LOC146336
ADXCRAG_BC028581_s_at	PIWIL1	9271	−0.72	Piwi-like 1 (*Drosophila*)
ADXCRAD_BX094012_s_at	SOX13	9580	−0.72	SRY (sex determining region Y)-box 13
ADXCRPDRC.4289.C1_at	RNF43	54894	1.38	Ring finger protein 43
ADXCRPD.10016.C1_at	SATB2	23314	1.82	SATB homeobox 2
ADXCRPDRC.8321.C1_s_at	TFCP2L1	29842	1.26	Transcription factor CP2-like 1
ADXCRIH.1549.C1_at	ELOVL5	60481	0.94	ELOVL family member 5, elongation of long chain fatty acids (FEN1/ Elo2, SUR4/Elo3-like, yeast)
ADXCRAG_BC028581_x_at	PIWIL1	9271	−1.72	Piwi-like 1 (*Drosophila*)
ADXCRIH.1305.C1_s_at	LYZ	4069	−1.61	Lysozyme
ADXCRSS.Hs#S1405714_at	RNF43	54894	1.27	Ring finger protein 43
ADXCRSS.Hs#S3740849_at	HSF5	124535	1.21	Heat shock transcription factor family member 5
ADXCRSS.Hs#S3012761_at	HSF5	124535	1.20	Heat shock transcription factor family member 5
ADXCRAD_BM825250_s_at	TM4SF4	7104	−0.99	Transmembrane 4 L six family member 4
ADXCRPD.7300.C1_s_at	LOC388199	388199	−1.28	Proline rich 25
ADXCRIH.4080.C1_s_at	SPINK1	6690	2.09	Serine peptidase inhibitor, Kazal type 1
ADXCRAD_NM_006113_s_at	VAV3	10451	1.38	Vav 3 guanine nucleotide exchange factor
ADXCRIH.546.C1_at	GGH	8836	1.49	γ-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)
ADXCRAD_AJ709424_s_at	ABLIM3	22885	−0.65	Actin binding LIM protein family, member 3
ADXCRPDRC.1943.C1_at	AXIN2	8313	1.32	Axin 2
ADXCRAD_BG470190_s_at	CDX2	1045	0.77	Caudal type homeobox 2
ADXCRAG_XM_371238_at	TRNP1	388610	−1.03	TMF1-regulated nuclear protein 1
ADXCRAD_BU664688_s_at	SLC14A1	6563	−0.82	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)
ADXCRPD.12823.C1_s_at	SYT13	57586	−0.77	Synaptotagmin XIII
ADXCRAD_CK823169_at	ANXA10	11199	−0.80	Annexin A10
ADXCRPD.8346.C1_at	HSF5	124535	1.34	Heat shock transcription factor family member 5
ADXCRPD.15182.C1_at	MIR142	406934	0.95	MicroRNA 142
ADXCRIH.31.C9_at	LYZ	4069	−1.61	Lysozyme
ADXCRAD_BP299698_s_at	VNN1	8876	−0.96	Vanin 1
ADXCRPD.14261.C1_at	ANO1	55107	−1.12	Anoctamin 1, calcium activated chloride channel
ADXCRAG_NM_002526_at	NT5E	4907	−1.27	5′-nucleotidase, ecto (CD73)
ADXCRAD_CN404528_s_at	DCBLD2	131566	−0.76	Discoidin, CUB and LCCL domain containing 2
ADXCRAD_BM852899_at	DUSP4	1846	−0.98	Dual specificity phosphatase 4
ADXCRAD_BP376354_at	AXIN2	8313	1.27	Axin 2
ADXCRAG_U04313_s_at	SERPINB5	5268	−0.89	Serpin peptidase inhibitor, clade B (ovalbumin), member 5
ADXCRIH.482.C1_at	KLK6	5653	−0.76	Kallikrein-related peptidase 6
ADXCRAD_BM718216_s_at	TRNP1	388610	−1.16	TMF1-regulated nuclear protein 1
ADXCRAG_XM_031357_s_at	KIAA0802	23255	−0.82	KIAA0802
ADXCRPD.1115.C1_s_at	MLPH	79083	−1.32	Melanophilin

NOTE. Positive LFC indicates higher expression in WT2.
Abbreviations: LFC, log fold change; WT2, double wild type.

denotes those predicted to be *BRAF* wild type. The pred-BRAFm samples consist of true *BRAF* mutants and the subset of WT2 and KRASm samples that are positive for the signature. These tumors share a common gene expression pattern, as can be seen in Appendix Figure A1 (online only). We call the subset of *BRAF* wild-type samples that are positive for the signature BRAFm-like to distinguish them from the true BRAFm.

Having identified a population of BRAFm-like samples, we proceeded to its characterization: In the population stratification analysis of PETACC-3, approximately 30% (76 of 257) of KRASm and 13%

Table 2. 32 Pairs of Genes Defining the *BRAF* Signature

Pair	Gene 1 (G1)	Gene 2 (G2)	Pair	Gene 1 (G1)	Gene 2 (G2)
1	C13orf18	CTSE	17	VAV3	OSBP2
2	DDC	AQP5	18	CFTR	KLK10
3	PPP1R14D	REG4	19	PHYH	DUSP4
4	HSF5	RSBN1L	20	PLCB4	HOXD3
5	SATB2	RASSF6	21	ZNF141	C11orf9
6	TNNC2	CRIP1	22	PPP1R14C	CD55
7	GGH	PPPDE2	23	FLJ32063	TRNP1
8	SPINK1	PLK2	24	APCDD1	FSCN1
9	PTPRO	TM4SF4	25	ACOX1	KIAA0802
10	ZSWIM1	MLPH	26	C10orf99	PLLP
11	RNF43	RBM8A	27	MIR142	IRX3
12	CELP	SOX8	28	ARID3A	SLC25A37
13	CBFA2T2	PIWIL1	29	C20orf111	PIK3AP1
14	PTPRD	LOC388199	30	AMACR	TPK1
15	CDX2	S100A16	31	AIFM3	ZIC2
16	TSPAN6	RBBP8	32	CTTNBP2	SERPINB5

NOTE. A sample is predicted to be *BRAF* mutant if the average expression of the genes in the Gene 1 (G1) columns is lower than the average expression of genes in Gene 2 (G2) columns.

(46 of 345) of WT2 samples were BRAFm-like. The BRAFm-like samples were significantly enriched in right-sided tumors in comparison with non–BRAF-like overall and also separately for KRASm (51% were right-sided) and WT2 (63% were right-sided). There was no association with a particular *KRAS* mutation subtype. Approximately 29% of the BRAFm-like samples were MSI-H (whereas 41% of the BRAFm were MSI-H). On the other hand, 50% of the MSI-H samples were BRAFm-like, with an additional 27% being BRAFm (Data Supplement). Separate hierarchical clustering of the KRASm and WT2 subpopulations, based on the genes from the signature, showed a split between BRAFm-like and the rest of the samples (Data Supplement). The identified BRAFm-like subpopulation was further described in terms of clinicopathologic features (Data Supplement), survival rates (Table 4 and Data Supplement), and differentially expressed genes between BRAFm-like and BRAFm samples (Data Supplement). The two groups of patients were similar with respect to their clinical and pathologic parameters, with the only exceptions being age (BRAFm-like comprise more patients older than 60 years) and tumor site (56% of BRAFm-like were right-sided, whereas 77% of BRAFm are right-sided; Data Supplement).

Table 3. Performance Metrics for the *BRAF* Signature

Data Set	Sensitivity	Specificity	Error Rate
PETACC-3[13]			
%	95.78	86.52	12.41
Standard deviation	4.04	0.18	0.14
Kim,[16] n = 20			
%	100.00	54.55	25.00
No.	9/9	6/11	5/20
Koinuma,[15] n = 20			
%	100.00	72.73	15.00
No.	9/9	8/11	3/20
Cetuximab,[17] n = 94			
%	85.71	91.95	8.51
No.	6/7	80/87	8/94
Aggregated, on validation sets, n = 134			
%	96.00	86.24	11.94
No.	24/25	94/109	16/134

NOTE. PETACC-3: cross-validation estimated performance. For the other data sets, the values indicate the observed performance.
Abbreviation: PETACC-3, Pan-European Trials in Alimentary Tract Cancers.

Prognostic Value of the Classifier

The prognostic value of the *BRAF* signature was assessed in the combined stage II and III population and in the stage III only subpopulation for three end points—overall survival (OS), relapse-free survival (RFS), and survival after relapse (SAR)—within the whole population, WT2 only, and KRASm only subpopulations, respectively. To account for the known prognostic effect of the MSI status (mainly for RFS) and its association with the *BRAF* mutation, the survival analysis was also performed within the MSS population only. The small number of MSI-H samples prevented a similar analysis of the signature predictions within MSI-H. In whole population and in MSS, the BRAFm and BRAFm-like patients have shorter survival times (OS and SAR), as can be seen in Figure 1 and the Data Supplement for different stratifications. The BRAFm-likeness showed the strongest prognostic effect for SAR, for both KRASm and WT2 (in all and MSS-only samples; see Figs 1F and 1H). The corresponding hazard ratios and their 95% CIs as well as the corresponding log-rank test *P* values for each of these comparisons are summarized in Table 4.

No statistically significant difference in survival was found between the BRAFm and BRAFm-like subpopulations, even though a tendency was observed for the patients with a BRAFm-like tumor to have a slightly better prognosis than those with a BRAFm tumor.

To identify potential drivers of the prognostic effect, we assessed the prognostic value of each of the 64 genes in the signature by fitting univariate Cox regression models in the whole PETACC-3 population and in the subset of *BRAF* wild-type samples (KRASm and WT2). Most of these genes were found to be significantly associated with the SAR end point, and, for 25 of them, the association was found also in the *BRAF* wild-type subgroup. These results reveal multiple interesting genes for future studies (Data Supplement).

External Validation

The *BRAF* signature was validated on three external data sets: Koinuma,[15] Kim,[16] and an internal series of patients with cetuximab-treated stage IV disease with gene expression data from primary tumors.[17] When genes from the signature were not represented on a platform, only the complete pairs of genes were considered. The aggregated observed sensitivity was 96.0% (24 of 25 BRAFm correctly identified) and the specificity was 86.24% (94 of 109 WT2 and KRASm correctly predicted; Table 3). This confirmed the highly sensitive recognition of tumors with a BRAFm and their distinction from majority non-BRAFm tumors, whereas approximately 14% of the latter were also wrongly classified as BRAFm. The reported specificity refers to KRASm and WT2 samples that should have been labeled as BRAF wild type by the classifier. The existence of a BRAFm-like group of patients is thus confirmed in these data sets.

The prognostic value of the *BRAF* signature has been validated in all and in the stage II and III only samples from the Moffitt data set[18] for OS and SAR (RFS being only marginally significant in stage II and III). No information on *BRAF* or *KRAS* mutational status was available,

Table 4. Survival Analyses Results

	OS			RFS			SAR		
Data Set	*P*	HR	95% CI	*P*	HR	95% CI	*P*	HR	95% CI
PETACC-3, all									
pred-BRAFm/*pred-BRAFwt*	**.0005**	**1.67**	**1.25 to 2.25**	.2447	1.17	0.90 to 1.53	**< .001**	**2.85**	**2.06 to 3.95**
BRAFm/*BRAFwt*	**.0021**	**2.01**	**1.28 to 3.17**	.1602	1.37	0.88 to 2.12	**< .001**	**3.68**	**2.20 to 6.16**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.5196	1.16	0.74 to 1.83	.4724	1.17	0.76 to 1.78	**.0021**	**2.13**	**1.30 to 3.48**
Within WT2: BRAFm-like/*pred-BRAFwt*	.1312	1.58	0.87 to 2.87	.4866	1.20	0.72 to 2.01	**.0011**	**2.72**	**1.46 to 5.06**
PETACC-3, stage III									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**1.93**	**1.41 to 1.79**	.0455	1.34	1.00 to 1.79	**< .0001**	**3.04**	**2.15 to 4.29**
BRAFm/*BRAFwt*	**.0024**	**2.14**	**1.29 to 3.55**	.1685	1.41	0.86 to 2.32	**< .0001**	**4.53**	**2.54 to 8.07**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.1916	1.37	0.85 to 2.21	.8203	1.05	0.68 to 1.64	**.0038**	**2.09**	**1.26 to 3.46**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0365	1.90	1.03 to 3.50	.2154	1.40	0.82 to 2.40	**.0012**	**2.75**	**1.45 to 5.19**
PETACC-3, MSS									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**2.19**	**1.57 to 3.07**	.0159	1.46	1.07 to 1.99	**< .0001**	**3.16**	**2.17 to 4.59**
BRAFm/*BRAFwt*	**< .0001**	**2.91**	**1.74 to 4.88**	.0228	1.79	1.08 to 2.98	**< .0001**	**4.67**	**2.57 to 8.45**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.0511	1.59	0.99 to 2.53	.4690	1.17	0.76 to 1.82	**.0043**	**2.07**	**1.24 to 3.43**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0642	1.98	0.95 to 4.16	.3464	1.37	0.71 to 2.63	**.0001**	**4.24**	**1.89 to 9.47**
PETACC-3, MSS/stage III									
pred-BRAFm/*pred-BRAFwt*	**< .0001**	**2.27**	**1.58 to 3.25**	.0105	1.54	1.10 to 2.15	**< .0001**	**2.97**	**2.01 to 4.40**
BRAFm/*BRAFwt*	**.0024**	**2.43**	**1.35 to 4.40**	.1149	1.59	0.89 to 2.86	**< .0001**	**3.88**	**1.99 to 7.56**
Within KRASm: BRAFm-like/*pred-BRAFwt*	.0216	1.77	1.08 to 2.89	.1765	1.37	0.87 to 2.16	**.0089**	**1.98**	**1.18 to 3.34**
Within WT2: BRAFm-like/*pred-BRAFwt*	.0220	2.35	1.11 to 4.98	.2789	1.46	0.73 to 2.93	**< .0001**	**4.67**	**2.05 to 10.63**
Moffitt[18]									
pred-BRAFm/*pred-BRAFwt*	.0376	1.67	1.02 to 2.73	.0956	1.77	0.90 to 3.50	**.0014**	**3.78**	**1.58 to 9.04**
pred-BRAFm/*pred-BRAFwt* (stages II,III)	**.0003**	**3.22**	**1.66 to 6.26**	.0498	2.02	0.99 to 4.15	**.0017**	**3.97**	**1.58 to 9.99**
pred-BRAFm/*pred-BRAFwt* (stage III)	**.0002**	**4.26**	**1.87 to 9.69**	.0204	2.79	1.13 to 6.87	**.0028**	**4.95**	**1.58 to 15.44**
	OS			PFS			SAR		
Cetuximab,[17] MSS	*P*	HR	95% CI	*P*	HR	95% CI	*P*	HR	95% CI
pred-BRAFm/*pred-BRAFwt*				**< .0001**	**4.49**	**2.40 to 8.38**	**< .0001**	**4.58**	**2.45 to 8.56**
BRAFm/*BRAFwt*				**.0018**	**3.24**	**1.46 to 7.19**	**< .0001**	**5.72**	**2.49 to 13.12**
Within BRAFwt: BRAFm-like/*pred-BRAFwt*				**.0017**	**3.45**	**1.56 to 7.63**	**< .0001**	**3.26**	**1.47 to 7.22**

NOTE. Highly significant results ($P < .01$) are set in bold. For the Cetuximab data set, only two end points could be considered: SAR and PFS. This data set contained also only stage IV MSS patients. When the predictions are considered within KRASm or WT2 subpopulations, those samples positive for the signature are called BRAFm-like (see the Results section). The comparison is given in the first column, with the reference category in italic font.

Abbreviations: BRAFm, true *BRAF* mutant; BRAFwt, true *BRAF* wild type; HR, hazard ratio; MSS, microsatellite stable; OS, overall survival; PETACC-3, Pan-European Trails in Alimentary Tract Cancers; PFS, progression-free survival; pred-BRAFm, classifier-predicted *BRAF* mutant; pred-BRAFwt, classifier-predicted *BRAF* wild type; SAR, survival after relapse.

making it impossible to draw any conclusions on the prognostic value of the signature within the KRASm or WT2 subpopulations. The signature was confirmed to be prognostic for SAR and progression-free survival (PFS) in the cetuximab[17] data set as well (OS information was not available for this data set). The survival analysis results and the corresponding Kaplan-Meier curves are given in Table 4 and in the Data Supplement.

DISCUSSION

Our results show that for c.1799T>A (p.V600E) BRAFm tumors, a characteristic gene expression signature of high sensitivity can be identified, and this signature extends to a population of *BRAF* wild-type subgroup of colon carcinomas (BRAFm-like) sharing similar clinicopathologic and gene expression features of potential prognostic importance. The *BRAF* mutation status has been previously shown to have prognostic value in CRC,[7,25-27] both in MSS and MSI-H tumors, and this feature is also shared by our signature in the case of MSS tumors. Because of the limited number of MSI-H tumors, we could not assess its prognostic value in those samples. The BRAFm-like tumors, either KRASm or double wild type, show a similar poor prognostic in all and MSS-only samples. This effect was also independent of tumor stage.

Globally, the group of BRAFm-like tumors discovered studying the gene expression data shows clinicopathologic features more similar to the BRAFm tumors (Data Supplement) than to pred-BRAFwt. As previously described,[13,28] BRAFm tumors are found with higher frequencies in right (proximal) colon, are enriched for the MSI-H phenotype, and are of higher grade. In our study, the frequencies of high-grade were 30% in BRAFm, 20% in BRAFm-like, and 5% in pred-BRAFwt; of MSI-H, 30%, 30%, and 3%, respectively; of right-side, 75%, 55%, and 30%, respectively. The mucinous tumors are most frequently BRAFm-like (45%) and are less often BRAFm (30% *v* only 10% in pred-BRAFwt). The exception is age, for which the frequency of young patients is highest in BRAFm-like (55%) and lowest in BRAFm (35%).

From a biologic perspective, this finding supports the notion that the poor outcome of tumors with BRAFm is shared with some non–BRAF-mutated tumors, suggesting that they have common biology that drives poor survival after relapse. For the genes in the signature, the c.1799T>A (p.V600E) BRAFm tumors display a homogeneous

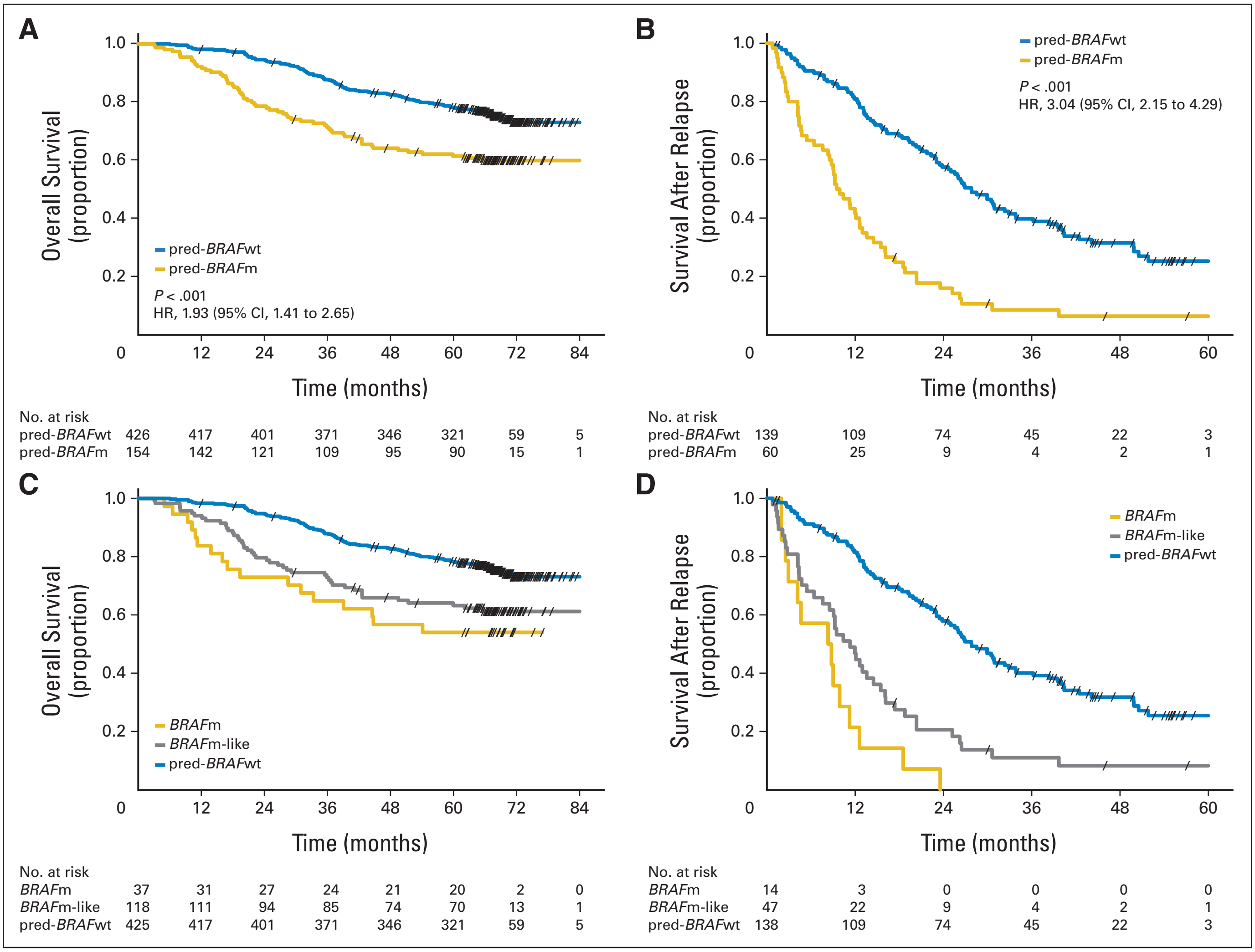


Fig 1. Kaplan-Meier curves for different stratifications of the stage III subpopulation and different end points. Columns correspond to overall survival and survival after relapse end points, respectively. Panels A-D correspond to stratifications into samples predicted to be *BRAF* mutant (pred-BRAFm)/predicted to be *BRAF* wild type (pred-BRAFwt; A, B) and *BRAF* mutant (BRAFm)/*BRAF* mutant like (BRAFm-like)/pred-BRAFwt (C, D) in the whole stage III subpopulation. Panels E-H correspond to stratifications BRAFm-like/pred-BRAFwt within *KRAS* mutant (E, F) and double wild type (WT2; G, H) subpopulations, in microsatellite stable. For the cases when only two populations are compared, the log-rank test *P* values and the hazard ratios (HRs; with 95% CIs) are given.

gene expression pattern, which is also found in some KRASm and WT2 samples (approximately 30% and 13% in our data, respectively; Appendix Fig A1). It is interesting to note that *BRAF* mutations have been strongly associated with the serrated adenoma pathway,[29,30] and thus the clear differences in gene expression between BRAFm and other colon tumors may be related to a different adenoma-carcinoma progression sequence. The existence of several subgroups of CCs, defined by their DNA methylation and mutation status, was first discovered in a population-based study[31] and was then subsequently confirmed.[32,33] A recent study[34] similarly presented evidence validating the existence of a cluster that included all BRAFm samples and a fraction of KRASm (18% of all KRASm) and WT2 samples and that was enriched for CIMP-positive, MLH1 hypermethylated, and right-sided tumors. For the moment, we can only speculate about the relation between our BRAFm-like concept and this cluster. In any case, it also supports the idea that c.1799T>A (p.V600E) BRAFm tumors form a homogeneous group with respect to the genes in the signature and that a sizeable set of other tumors show similar characteristics. The underlying driver biology of this BRAFm-like group remains unknown, although it is clearly associated with clinicopathologic features, such as MSI-H, right-sidedness, and mucinous histology.

The identification of a BRAFm-like subpopulation of CC that includes KRASm and WT2 samples and that manifests a coherent clinical behavior suggests that a new definition of CC subgroups is needed. To the best of our knowledge, this is the first reported split based on gene expression data of the KRASm tumors (see also Data Supplement), which were considered until now as a compact group, based solely on their mutation status.

The genes associated with the *BRAF* c.1799T>A (p.V600E) mutation in CC and in melanoma are dissimilar, indicating tissue-specific biology that needs to be understood and targeted differently. It is therefore not surprising that *BRAF*-specific inhibitors, such as PLX4032 or GSK2118436, although very successful in BRAFm melanoma, have failed in BRAFm colorectal cancer treatment.[35,36]

In summary, our results show that for c.1799T>A (p.V600E) BRAFm tumors, a high-sensitivity gene expression signature can be

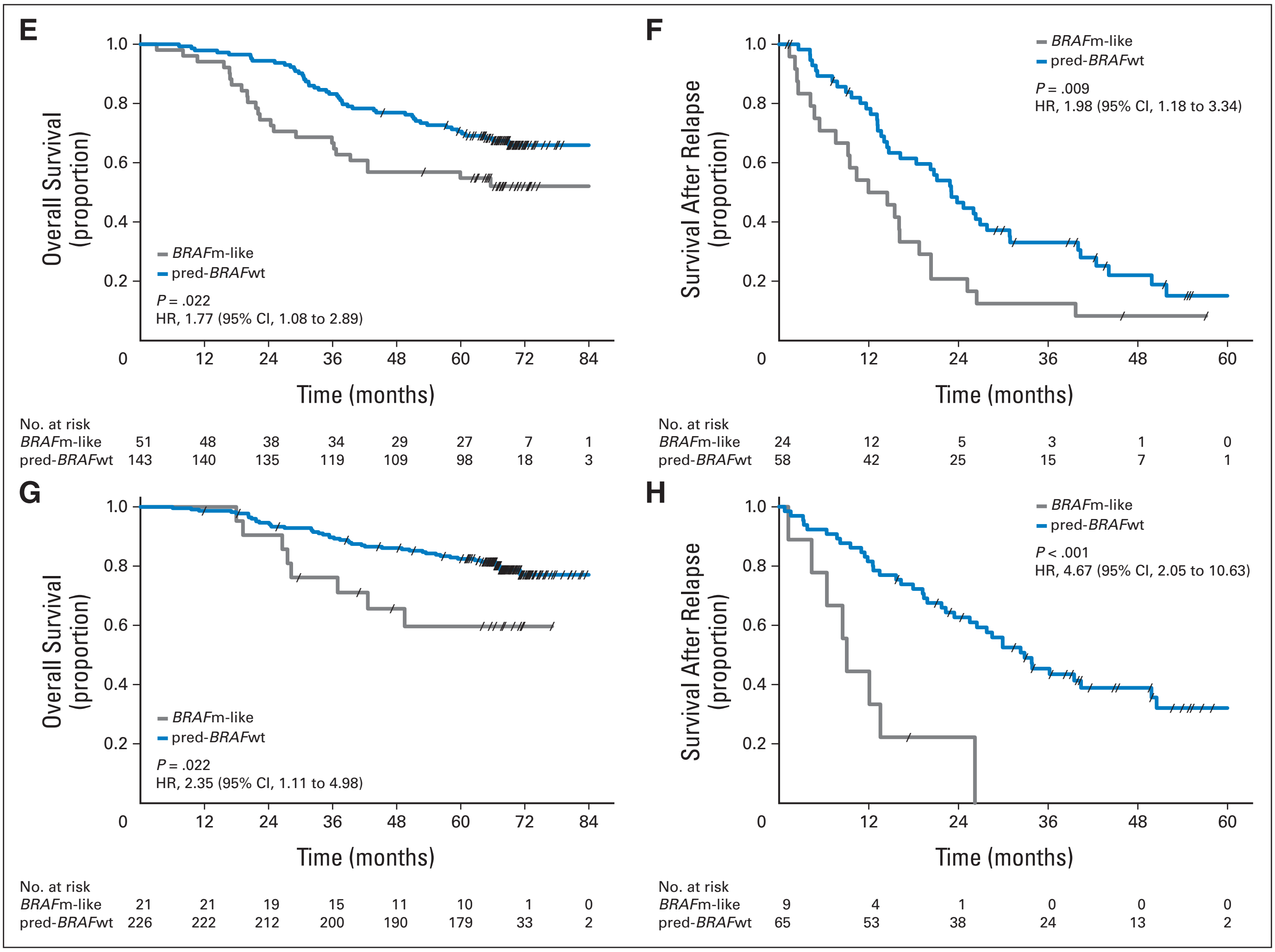


Fig 1. (continued).

derived and that this signature identifies also a subgroup of BRAFm-like tumors sharing similar clinicopathologic features of potential prognostic importance. They also indicate histologic and prognostic heterogeneity within the KRASm and thus challenge the current assumption that these tumors can all be considered alike. This stratification may be of interest in randomized clinical trials and in drug development studies and can easily be obtained by applying the proposed classifier.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: Scott Weinrich, Pfizer (C); Heather Estrella, Pfizer (C); Graeme Hodgson, Pfizer (C); Tao Xie, Pfizer (C) **Consultant or Advisory Role:** Sabine Tejpar, Pfizer (C); Fred T. Bosman, Pfizer (C); Arnaud D. Roth, Pfizer (C) **Stock Ownership:** Scott Weinrich, Pfizer; Heather Estrella, Pfizer; Graeme Hodgson, Pfizer **Honoraria:** Arnaud D. Roth, Pfizer **Research Funding:** Sabine Tejpar, Pfizer, Merck; Eric Van Cutsem, Pfizer; Fred T. Bosman, Pfizer **Expert Testimony:** None **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Vlad Popovici, Eva Budinska, Sabine Tejpar, Arnaud D. Roth, Mauro Delorenzi
Provision of study materials or patients: Eric Van Cutsem
Collection and assembly of data: Vlad Popovici, Eva Budinska, Sabine Tejpar, Scott Weinrich, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Tao Xie, Fred T. Bosman, Arnaud D. Roth
Data analysis and interpretation: Vlad Popovici, Eva Budinska, Sabine Tejpar, Heather Estrella, Graeme Hodgson, Eric Van Cutsem, Fred T. Bosman, Mauro Delorenzi
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Samowitz WS, Albertsen H, Herrick J, et al: Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology 129:837-845, 2005

2. Nosho K, Irahara N, Shima K, et al: Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. PLoS One 3:e3698, 2008

3. Brink M, de Goeij AF, Weijenberg MP, et al: K-ras oncogene mutations in sporadic colorectal cancer in The Netherlands Cohort Study. Carcinogenesis 24:703-710, 2003

4. English DR, Young JP, Simpson JA, et al: Ethnicity and risk for colorectal cancers showing somatic BRAF V600E mutation or CpG island methylator phenotype. Cancer Epidemiol Biomarkers Prev 17:1774-1780, 2008

5. Rosenberg DW, Yang S, Pleau DC, et al: Mutations in BRAF and KRAS differentially distinguish serrated versus non-serrated hyperplastic aberrant crypt foci in humans. Cancer Res 67:3551-3554, 2007

6. Velho S, Moutinho C, Cirnes L, et al: BRAF, KRAS and PIK3CA mutations in colorectal serrated polyps and cancer: Primary or secondary genetic events in colorectal carcinogenesis? BMC Cancer 8:255, 2008

7. Roth AD, Tejpar S, Delorenzi M, et al: Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: Results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. J Clin Oncol 28:466-474, 2010

8. Pratilas CA, Xing F, Solit DB: Targeting oncogenic BRAF in human cancer. Curr Top Microbiol Immunol [epub ahead of print on August 5, 2011]

9. Pratilas CA, Taylor BS, Ye Q, et al: (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. Proc Natl Acad Sci U S A 106:4519-4524, 2009

10. Dry JR, Pavey S, Pratilas CA, et al: Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). Cancer Res 70:2264-2273, 2010

11. Pavey S, Johansson P, Packer L, et al: Microarray expression profiling in melanoma reveals a BRAF mutation signature. Oncogene 23:4060-4067, 2004

12. Kannengiesser C, Spatz A, Michiels S, et al: Gene expression signature associated with BRAF mutations in human primary cutaneous melanomas. Mol Oncol 1:425-430, 2008

13. Van Cutsem E, Labianca R, Bodoky G, et al: Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. J Clin Oncol 27:3117-3125, 2009

14. De Roock W, Claes B, Bernasconi D, et al: Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: A retrospective consortium analysis. Lancet Oncol 11:753-762, 2010

15. Koinuma K, Yamashita Y, Liu W, et al: Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. Oncogene 25:139-146, 2006

16. Kim IJ, Kang HC, Jang SG, et al: Oligonucleotide microarray analysis of distinct gene expression patterns in colorectal cancer tissues harboring BRAF and K-ras mutations. Carcinogenesis 27:392-404, 2006

17. Budinska E, Delorenzi M, De Roock W, et al: New insights to gene expression signatures from primary FFPE tumors for the prediction of response to cetuximab in KRAS and BRAF wild-type colorectal cancer (CRC). J Clin Oncol 28, 243s, 2010 (suppl; abstr 3588)

18. Smith JJ, Deane NG, Wu F, et al: Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138:958-968, 2010

19. Irizarry RA, Bolstad BM, Collin F, et al: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31:e15, 2003

20. Smyth GK: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3, 2004

21. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57:289-300, 1995

22. Tan AC, Naiman DQ, Xu L, et al: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21:3896-3904, 2005

23. Matthews BW: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442-451, 1975

24. Shi L, Campbell G, Jones WD, et al: The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827-838, 2010

25. Samowitz WS, Sweeney C, Herrick J, et al: Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. Cancer Res 65:6063-6069, 2005

26. Ogino S, Nosho K, Kirkner GJ, et al: CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. Gut 58:90-96, 2009

27. French AJ, Sargent DJ, Burgart LJ, et al: Prognostic significance of defective mismatch repair and BRAF V600E in patients with colon cancer. Clin Cancer Res 14:3408-3415, 2008

28. Li WQ, Kawakami K, Ruszkiewicz A, et al: BRAF mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status. Mol Cancer 5:2, 2006

29. Snover DC: Update on the serrated pathway to colorectal carcinoma. Hum Pathol 42:1-10, 2011

30. Leggett B, Whitehall V: Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology 138:2088-2100, 2010

31. Ogino S, Kawasaki T, Kirkner GJ, et al: CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: Possible associations with male sex and KRAS mutations. J Mol Diagn 8:582-588, 2006

32. Yagi K, Akagi K, Hayashi H, et al: Three DNA methylation epigenotypes in human colorectal cancer. Clin Cancer Res 16:21-33, 2010

33. Dahlin AM, Palmqvist R, Henriksson ML, et al: The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. Clin Cancer Res 16:1845-1855, 2010

34. Hinoue T, Weisenberger DJ, Lange CP, et al: Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res [epub ahead of print on June 9, 2011]

35. Kopetz S, Desai J, Chan E, et al: PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. J Clin Oncol 28:269s, 2010 (suppl; abstr 3534)

36. Kefford R, Arkenau H, Brown MP, et al: Phase I/II study of GSK2118436, a selective inhibitor of oncogenic mutant BRAF kinase, in patients with metastatic melanoma and other solid tumors. J Clin Oncol 28:611s, 2010 (suppl; abstr 8503)

[*8*] Popovici V, **Budinska E,** Bosman FT, Tejpar S, Roth AD, Delorenzi M. Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer. *BMC Cancer.* 2013 Sep 12;13:439. doi:10.1186/1471-2407-13-439.

BMC Cancer

RESEARCH ARTICLE Open Access

Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer

Vlad Popovici[1,2*], Eva Budinska[1,2], Fred T Bosman[3], Sabine Tejpar[4], Arnaud D Roth[5] and Mauro Delorenzi[2,6]

Abstract

Background: The mutation status of the BRAF and KRAS genes has been proposed as prognostic biomarker in colorectal cancer. Of them, only the BRAF V600E mutation has been validated independently as prognostic for overall survival and survival after relapse, while the prognostic value of KRAS mutation is still unclear. We investigated the prognostic value of BRAF and KRAS mutations in various contexts defined by stratifications of the patient population.

Methods: We retrospectively analyzed a cohort of patients with stage II and III colorectal cancer from the PETACC-3 clinical trial (N = 1,423), by assessing the prognostic value of the BRAF and KRAS mutations in subpopulations defined by all possible combinations of the following clinico-pathological variables: T stage, N stage, tumor site, tumor grade and microsatellite instability status. In each such subpopulation, the prognostic value was assessed by log rank test for three endpoints: overall survival, relapse-free survival, and survival after relapse. The significance level was set to 0.01 for Bonferroni-adjusted p-values, and a second threshold for a trend towards statistical significance was set at 0.05 for unadjusted p-values. The significance of the interactions was tested by Wald test, with significance level of 0.05.

Results: In stage II-III colorectal cancer, BRAF mutation was confirmed a marker of poor survival only in subpopulations involving microsatellite stable and left-sided tumors, with higher effects than in the whole population. There was no evidence for prognostic value in microsatellite instable or right-sided tumor groups. We found that BRAF was also prognostic for relapse-free survival in some subpopulations. We found no evidence that KRAS mutations had prognostic value, although a trend was observed in some stratifications. We also show evidence of heterogeneity in survival of patients with BRAF V600E mutation.

Conclusions: The BRAF mutation represents an additional risk factor only in some subpopulations of colorectal cancers, in others having limited prognostic value. However, in the subpopulations where it is prognostic, it represents a marker of much higher risk than previously considered. KRAS mutation status does not seem to represent a strong prognostic variable.

Keywords: Colorectal cancer, BRAF V600E mutation, KRAS mutations, Survival analysis, Stratified analysis

* Correspondence: popovici@iba.muni.cz
[1]Institute of Biostatistics and Analyses, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic
[2]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Lausanne, Switzerland
Full list of author information is available at the end of the article

Background

Our current models of colorectal cancer (CRC) are dominated by the idea of a sequential tumor progression from adenoma to carcinoma, in which the accumulation of genetic events in key genes defines alternative oncogenic paths with impact on tumor characteristics. These genetic events include the mutational activation of oncogenes like BRAF and KRAS, disruption of WNT signaling, allelic imbalance on chromosome 18q and mutation of TP53 tumor suppressor gene [1-4]. Since the mutations of BRAF and KRAS genes, which lead to the activation of MEK/ERK pathway, are seen as important events in the tumor progression and based on their relatively high incidence (7-15% for BRAF mutations and 35-40% for KRAS mutations [5-8]), they have been proposed as prognostic biomarkers for CRC. Of them, only BRAF V600E mutation has been consistently validated, while the prognostic value of KRAS mutation remains debatable. The BRAF has been shown to be prognostic for overall survival (OS) and survival after relapse (SAR) in general CRC population by us and others [9-13] as well as in microsatellite-stable (MSS) population [12,14], while having no prognostic value for relapse-free survival (RFS). In these studies, the hazard ratios (HR) for BRAF mutation varied between 1.4 and 2.1 for OS and 2.3 to 3.6 for SAR. In the case of KRAS mutation, the published results are contradictory, with prognostic value, in the positive studies, found only for relapse-free survival [9,11,15], while other studies, including our own [13], did not find any evidence of prognostic value for KRAS mutation. Also, a recent meta-analytical review found no evidence supporting the prognostic value of KRAS mutation [16]. A detailed review is given in [17].

The question remains whether the prognostic value of the BRAF and KRAS mutations is uniform across different patient groups defined by clinical parameters or if there are interactions that would influence their utility. Taking advantage of a large series of stage II-III CRC tumors with mutation data from the PETACC-3 clinical trial [18], we systematically investigate the prognostic value of the BRAF and KRAS mutations in all possible stratifications – contexts – defined by a set of clinical parameters found to be important in survival prognosis in a previous analysis [19]. The main question our study tries to answer is whether the mutations of BRAF and KRAS genes are indicators of different prognosis within otherwise uniform (with respect to the clinical parameters considered) subpopulations of patients with CRC. A secondary question we address, for the main findings, is whether the observed prognostic values are statistically significant also in multivariate models, in the respective subpopulations.

Methods

We retrospectively analyzed the PETACC-3 clinical trial [18] data set (N = 1,423), of patients with stage II and III CRC, by generating the subpopulations defined by all possible combinations of levels of the following five variables: MSI status (MSI-H and MSS levels), tumor site (left and right), T stage (T1,2, T3, and T4), N stage (N0, N1 and N2) and tumor grade (G1,2 and G3,4). In total, there were 393 possible subpopulations (see Additional file 1 for an exhaustive listing), of which only those with more than N = 20 samples were further considered for testing the prognostic value of the BRAF and KRAS mutations. The full description of the data set is given in [19].

In each subpopulation, the prognostic importance of the BRAF and KRAS mutations was assessed using log-rank test comparing the survival of BRAF-/KRAS-mutant population to the BRAF- and KRAS- wild type (double wild type – WT2) population, for overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR) endpoints. Data was summarized with hazard ratios (HR), their 95% confidence intervals (CI), P-values and adjusted P-values (Bonferonni correction, denoted hereinafter by P^*). For a result to be considered statistically significant we required that $P^* \leq 0.01$ and that at least 10 patients were in each of the two groups compared. If only $P \leq 0.05$, the result was reported as a trend towards significance. The significance of the interactions was tested by Wald test in the presence of both main effects, with significance level of 0.05 (no adjustment for multiple testing in this case). All tests were two-sided.

All computations were carried out in R version 2.15.2 (http://www.r-project.org) and survival analysis was performed using R survival package version 2.37-2.

Results and discussion

In the global population, the BRAF mutation is prognostic for poorer overall survival and survival after relapse, while KRAS mutation is not prognostic for any of the three endpoints (Table 1). In stratified analyses and after correction for multiple testing, BRAF mutation status remained a significant prognostic marker in various subpopulations. On the contrary, KRAS mutation status never reached the level of significance required after P-value adjustment ($P^* \leq 0.01$ and at least 10 patients in both of the groups compared). However, in several stratifications, KRAS mutation showed a trend towards significance ($P \leq 0.05$). The full table of results with all possible stratifications is given as Additional file 1.

BRAF mutation

The BRAF mutation was prognostic for overall survival in MSS and/or left-sided tumors subpopulations (Figure 1). In the MSS tumors, BRAF was indicative of worse overall survival ($P^* < 0.0001$; HR = 2.82; 95% CI = 1.85 to 4.30), as well as in MSS/left tumors ($P^* < 0.0001$; HR = 6.41; 95% CI = 3.57 to 11.52) and all left-sided tumors ($P^* < 0.0001$; HR = 5.18; 95% CI = 3.00 to 8.94) (Figure 1A,B). At the

Table 1 Univariate analysis of the prognostic factors in the whole CRC population

		OS		RFS		SAR	
Factor	**Comparison**	**P-value**	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**	**HR (95% CI)**
MSI	MSI-H vs MSS	0.0002	0.45 (0.30,0.69)	< 0.0001	0.48 (0.34,0.68)	0.9643	0.99 (0.65,1.52)
Site	Left vs Right	0.3143	0.89 (0.72,1.11)	0.2123	1.13 (0.93,1.36)	<0.0001	0.59 (0.47, 0.73)
Grade	G3,4 vs G1,2	0.0018	1.63 (1.29,2.23)	0.0012	1.56 (1.19,2.04)	0.0387	1.38 (1.01,1.88)
T stage	T3 vs T1,2	0.0634	1.76 (0.96,3.22)	0.0629	1.58 (0.97,2.58)	0.1399	1.57 (0.86,2.88)
	T4 vs T1,2	0.0002	3.06 (1.63,5.74)	< 0.0001	2.69 (1.61,4.48)	0.0680	1.78 (0.95,3.35)
N stage	N1 vs N0	< 0.0001	1.91 (1.38,2.65)	< 0.0001	1.78 (1.36, 2.32)	0.9809	0.98 (0.71,1.35)
	N2 vs N0	< 0.0001	4.51 (3.28,6.21)	< 0.0001	4.06 (3.11,5.29)	0.1498	1.24 (0.90,1.71)
BRAF	BRAF mut vs WT2	0.0004	1.92 (1.33,2.78)	0.0832	1.35 (0.96,1.89)	< 0.0001	2.56 (1.75,3.70)
	BRAF mut vs BRAF wt	0.0009	1.78 (1.26,2.53)	0.1174	1.30 (0.94,1.81)	< 0.0001	2.48 (1.74,3.53)
KRAS	KRAS mut vs WT2	0.1461	1.20 (0.93,1.54)	0.4410	1.09 (0.88,1.33)	0.1755	1.18 (0.93,1.52)
	KRAS mut vs KRAS wt	0.4826	1.09 (0.86,1.37)	0.7245	1.04 (0.85,1.27)	0.7222	1.04 (0.82,1.32)

same time, BRAF mutation was not prognostic in any stratification involving only right-sided tumors (Figure 1C) and/or MSI-H tumors. In a multivariate model, including up to second degree interactions between MSI status, BRAF mutation and tumor site, adjusted for grade, T stage and N stage, the only significant interaction was between BRAF mutation and tumor site (P = 0.0041). The interaction between BRAF mutation status and tumor site was also significant within MSS tumors (P = 0.0033), but not within MSI-H tumors. The interaction between BRAF mutation status and MSI status was not significant in either left or right-sided tumors. These results show that BRAF mutation represents an additional risk factor only within MSS/left tumors, with no statistically significant effect in right or MSI-H tumors, the general prognostic value of BRAF mutation being driven by its effect in this subpopulation. As a consequence, the corresponding HR should be re-interpreted: a BRAF mutation does not double the risk of death for all patients carrying this mutation (HR = 1.92 in global population), but represents a six-fold increase of the risk in the case of patients with MSS/left tumors (HR = 6.41) – in comparison with the double wild type MSS/left tumors. At the same time, BRAF mutation does not significantly influence the risk of death (in comparison with WT2) in MSI-H and/or right-sided tumors. The MSS/left side BRAF-mutant population emerges as the worst surviving group of patients in our data set: for example, the 3-year overall survival rate is 0.35 (95% CI = 0.20 to 0.66) in comparison to 0.89 (95% CI = 0.85 to 0.93) for KRAS-mutant and 0.91 (95% CI = 0.88 to 0.93) for WT2, respectively (Table 2). The observation could not be extended to MSS/right-sided tumors (Table 2).

Interestingly, BRAF mutation was also prognostic for shorter relapse-free survival in left-sided tumors (Figure 2): all left-sided tumors ($P^* = 0.0002$; HR = 3.31; 95% CI = 1.98 to 5.55) and MSS/left tumors ($P^* = 0.0005$; HR = 3.57; 95% CI = 2.02 to 6.31) (Figure 2, see also Table 2). This

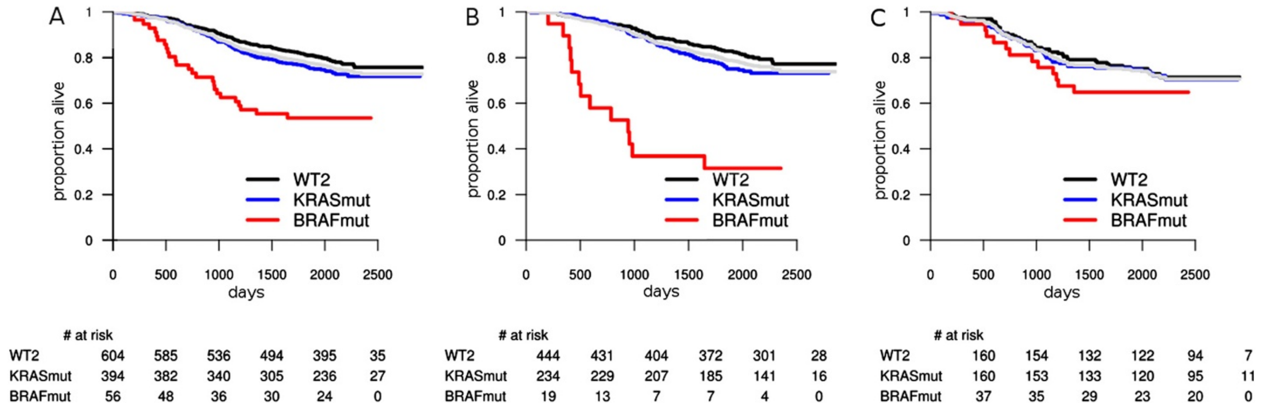


Figure 1 Overall survival: prognostic value of BRAF and KRAS mutations within MSS and by tumor site. A: all MSS tumors; **B**: MSS left-sided tumors; **C**: MSS right-sided tumors. The light gray survival curve represents the whole subpopulation survival (**A**: all MSS, **B**: MSS left-sided, **C**: MSS right-sided tumors).

Table 2 Three-year overall and relapse-free survival rates, and one-year survival after relapse rates in MSS/left and MSS/right populations, stratified by mutation status

	MSS/left		MSS/right	
Population	Survival rate	95% CI	Survival rate	95% CI
	OS: 3-year survival rates			
WT2	0.91	0.88-0.93	0.83	0.77-0.89
KRAS mut	0.89	0.85-0.93	0.80	0.74-0.86
BRAF mut	0.37	0.20-0.66	0.73	0.60-0.89
	RFS: 3-year survival rates			
WT2	0.75	0.71-0.80	0.75	0.68-0.82
KRAS mut	0.68	0.62-0.74	0.73	0.66-0.80
BRAF mut	0.32	0.16-0.61	0.68	0.54-0.84
	SAR: 1-year survival rates			
WT2	0.81	0.74-0.88	0.65	0.52-0.82
KRAS mut	0.80	0.71-0.89	0.75	0.53-0.80
BRAF mut	0.17	0.05-0.60	0.36	0.17-0.79

is a novel observation, since BRAF mutation was not generally considered prognostic for relapse. In other MSS-subpopulations involving left-sided tumors BRAF mutation is also prognostic (see Additional file 1). Again, the BRAF mutation was not prognostic in any subpopulation involving MSI-H and/or right-sided tumors. In a multivariate model, involving up to second degree interactions between MSI status, BRAF mutation and tumor site, adjusted for grade, T stage and N stage, the only significant interaction was between BRAF mutation and tumor site (P = 0.047). The interaction between BRAF mutation status and tumor site was also significant within MSS tumors (P = 0.043), but not within MSI-H tumors (where the small number of BRAF mutants in the left colon limits the statistical power). Hence, the prognostic value of the BRAF mutation is confined to the MSS/left-sided tumors.

For the survival after relapse (SAR), BRAF mutation represents an additional risk factor in more stratifications, most of them involving MSS and/or left-sided tumors. BRAF mutation shows also a trend to be prognostic in MSS/right-sided tumors as well, even though the p-value was no longer significant after multiple testing correction. The BRAF mutation was indicative of poor survival after relapse in all MSS tumors ($P^* < 0.0001$; HR = 3.43; 95% CI = 2.19 to 5.36); MSS/left tumors ($P^* = 0.0002$; HR = 3.89; 95% CI = 2.11 to 7.20) and showed a trend in MSS/right (P = 0.0111; HR = 2.27; 95% CI = 1.17 to 4.38) (Figure 3). The test for interaction between BRAF mutation status and tumor site was not significant, hence we conclude that BRAF mutation is prognostic for SAR in all MSS patients.

The differences in prognostic value of the BRAF mutation status in various subpopulations suggest a certain degree of heterogeneity in the survival of patients harboring this mutation. Indeed, within the BRAF mutant population, the MSS tumors had worse outcome for overall survival (P = 0.0021; HR = 3.45; 95% CI = 1.49 to 7.69)) and relapse-free survival (P = 0.0085; HR = 2.63; 95% CI = 1.25 to 5.56), this observation being in line with the fact that MSI-H has a protective prognostic effect in CRC. At the same time, the left BRAF-mutant tumors had a worse prognosis than the right BRAF-mutant tumors, for overall survival (within all BRAF-mutants: P = 0.0003; HR = 3.20; 95% CI = 1.64 to 6.23; within MSS/BRAF-mutants: P = 0.0059; HR = 2.84; 95% CI = 1.31 to 6.15; while within MSI-H/BRAF-mutants it could not be assessed) and for relapse-free survival (within all BRAF-mutants: P = 0.0002; HR = 3.24; 95% CI = 1.71 to 6.16; within MSS/BRAF-mutants: P = 0.0062; HR = 2.82; 95%

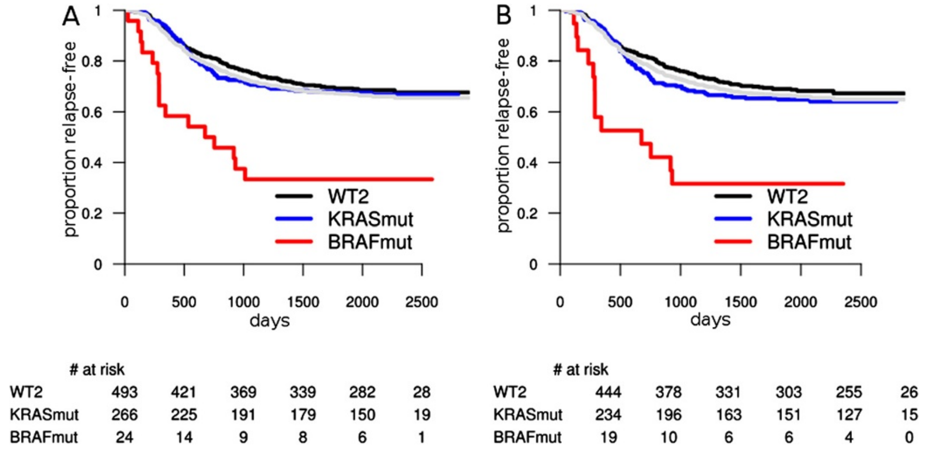
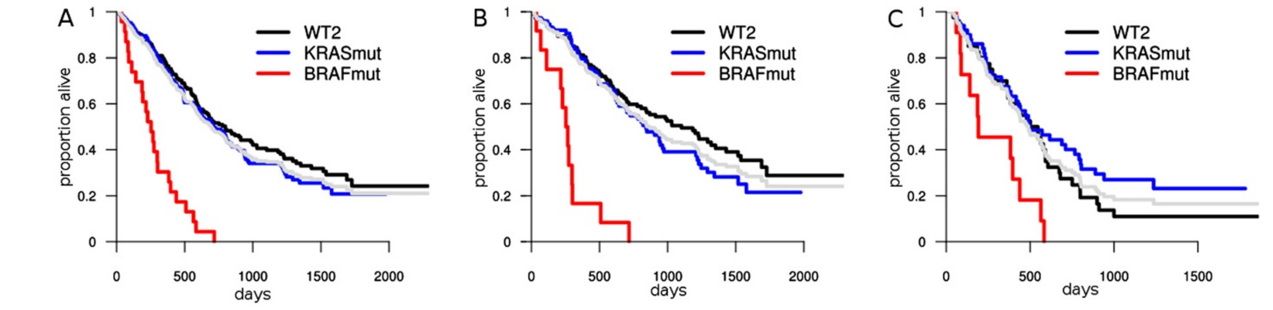


Figure 2 Relapse-free survival: prognostic value of BRAF and KRAS in left-sided tumors. A: all left-sided tumors; **B**: MSS left-sided tumors. The light gray survival curve represents the whole subpopulation survival (**A**: all left tumors; **B**: MSS left).



at risk

WT2	163	108	58	26	2
KRASmut	126	73	36	12	0
BRAFmut	23	4	0	0	0

at risk

WT2	123	87	54	24	2
KRASmut	75	50	27	9	0
BRAFmut	12	2	0	0	0

at risk

WT2	40	21	4	2
KRASmut	51	23	9	3
BRAFmut	11	2	0	0

Figure 3 Survival after relapse: prognostic value of BRAF and KRAS mutations in MSS tumors by site. A: all MSS tumors; **B**: MSS left-sided tumors; **C**: MSS right-sided tumors. The light gray survival curve represents the whole subpopulation survival (**A**: all MSS, **B**: MSS left-sided, **C**: MSS right-sided tumors).

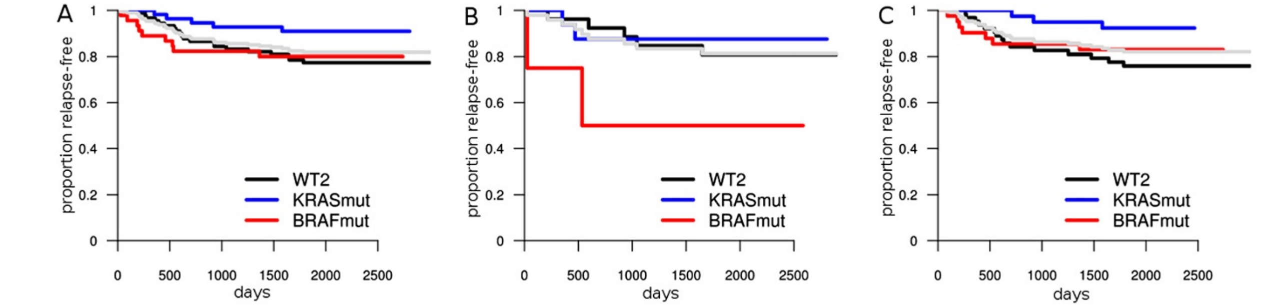
CI = 1.30 to 6.12; while within MSI-H/BRAF-mutants it could not be assessed). However, there was no statistically significant difference in survival after relapse among BRAF mutants, all having an equally poor survival.

KRAS mutation

KRAS mutation did not reach the significance level required to be considered prognostic for any of the three endpoints, since the adjusted p-values were all larger than 0.01. However, in some cases, it showed a trend towards significance ($P \leq 0.05$).

In overall survival, KRAS mutation had a trend to become significant in several stratifications of tumors with early stage lymph node invasion (N1). In all these, KRAS mutation was a marker of worse outcome (see Additional file 1). While not being a significant prognostic factor (as required by us) for relapse-free survival, KRAS mutation showed a trend to become prognostic. In contrast with BRAF, KRAS mutation seemed to be prognostic for RFS mostly in the right colon. The most intriguing observation was in MSI-H/right colon subpopulation (N = 102, KRAS mutants: 39), where KRAS mutation seemed to identify a low risk group (P = 0.0349; HR = 0.29; 95% CI = 0.08 to 0.99) (Figure 4). KRAS mutation was not prognostic for SAR. Also, no significant interaction between KRAS mutation, MSI status and tumor site was observed, for any of the three endpoints.

Since several studies have suggested that KRAS mutations at codon 12 may have a different prognostic value than codon 13 mutations [20], we have tested for differences in survival between the two groups of mutations, in all the same stratifications. No statistically significant difference was observed, but the sample size of our data might be too limited to detect such differences.



at risk

WT2	89	83	75	70	54	5
KRASmut	55	53	51	51	41	3
BRAFmut	45	39	37	36	31	6

at risk

WT2	26	25	23	22	16	1
KRASmut	16	14	14	14	13	3
BRAFmut	4	3	2	2	2	1

at risk

WT2	63	58	52	48	38	4
KRASmut	39	39	37	37	28	0
BRAFmut	41	36	35	34	29	5

Figure 4 Relapse-free survival: prognostic value of BRAF and KRAS in MSI-H tumors by site. A: all MSI-H tumors; **B**: MSI-H left-sided tumors; **C**: MSI-H right-sided tumors. The light gray survival curve represents the whole subpopulation survival (**A**: all MSI-H, **B**: MSI-H left-sided, **C**: MSI-H right-sided tumors).

Conclusions

In our analyses, we have compared the survival of BRAF/KRAS-mutated population with that of the double-wild type population, while controlling for several other parameters (tumor site, T and N stage, grade and MSI status).

Our analyses confirm the prognostic value of BRAF mutation status, in various stratifications. As a novelty, we observe a strong prognostic value for relapse-free survival of the BRAF mutation status in the MSS/left-colon tumors.

The interpretation of BRAF mutation as additional risk factor has to be made in the context of MSI status and tumor location. Indeed, our results show that BRAF represents a risk factor in the left colon and/or MSS tumors. In the data analyzed, we found no sufficient statistical evidence supporting a worse outcome associated with BRAF mutation in MSI-H tumors. As a consequence, the published hazard ratios for BRAF mutation for general population have to be reconsidered. The tumor staging (T or N stage, tumor grade) had a lesser impact on the prognostic value of the BRAF mutation status, while the tumor background (site and microsatellite (in)stability) significantly influenced the prognostic.

For the KRAS mutation, we could not confirm nor completely disprove its prognostic value. It was prognostic in several stratifications, in some showing a protective effect, while in others representing a risk factor. This is probably an effect of the heterogeneity of KRAS mutant population [21,22] and may explain in part the contradictory results published so far. With the strict requirements for statistical significance imposed, KRAS mutation did not appear to have prognostic value in any of the stratifications. The trend towards significance suggests, however, a potential utility as prognostic marker for RFS mostly in right colon.

In conclusion, the utility of the BRAF and KRAS as prognostic biomarkers depends on the MSI status and tumor location. We hypothesize that this interaction may extend to other biomarkers and prognostic gene signatures as well. At the same time, this observation has clear implications in clinical trial design and needs to be accounted for.

We make public the full table with all stratifications to support similar analyses in other data sets.

Additional file

Additional file 1: Full survival analysis results. In each possible stratification three endpoints were tested - overall survival, relapse-free survival and survival after relapse - and the sample size of the analysis along with the resulting p-values and hazard ratios are given.

Abbreviations

MSI: Microsatellite instability; MSI-H: High microsatellite instability; MSS: Microsatellite stable; OS: Overall survival; RFS: Relapse-free survival; SAR: Survival after relapse; WT2: Double wild type tumors: tumors that are BRAF- and KRAS-wild type.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VP conceived and designed the investigation, performed the statistical analyses, interpreted the results and drafted the manuscript. EB performed statistical analyses and drafted the manuscript. FB, ST, AR and MD participated in data generation, result interpretation and contributed to manuscript drafting. All authors have read and approved the final manuscript.

Acknowledgments

ST is a Senior Clinical Investigator of the Fund for Scientific Research Flanders (FWO-Vlaanderen) and has received research grants from the Belgian Federation against Cancer (Stichting tegen Kanker) and from the Belgian National Cancer Plan (Nationaal Kankerplan). This work was supported by the grants SNF 320030_135421 of the Swiss National Science Foundation and KFS 0269708-2010 of the Krebsliga Schweiz to AR and MD and by the EU Seventh Framework Programme, grant agreement 259015 to MD.

Author details

[1]Institute of Biostatistics and Analyses, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic. [2]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Institute of Pathology, Lausanne University Medical Center, Lausanne, Switzerland. [4]University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Leuven, Belgium. [5]Oncosurgery Unit, Geneva University Hospital, Geneva, Switzerland. [6]University of Lausanne, Lausanne, Switzerland.

Received: 13 May 2013 Accepted: 24 September 2013
Published: 27 September 2013

References

1. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL: **Genetic alterations during colorectal-tumor development.** *N Engl J Med* 1988, **319:**525–532.
2. Vogelstein B, Fearon ER, Kern SE, Hamilton SR, Preisinger AC, Nakamura Y, White R: **Allelotype of colorectal carcinomas.** *Science* 1989, **244:**207–211.
3. Markowitz SD, Bertagnolli MM: **Molecular origins of cancer: Molecular basis of colorectal cancer.** *N Engl J Med* 2009, **361:**2449–2460.
4. Segditsas S, Tomlinson I: **Colorectal cancer and genetic alterations in the Wnt pathway.** *Oncogene* 2006, **25:**7531–7537.
5. Samowitz WS, Sweeney C, Herrick J, Albertsen H, Levin TR, Murtaugh MA, Wolff RK, Slattery ML: **Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers.** *Cancer Res* 2005, **65:**6063–6069.
6. Nosho K, Irahara N, Shima K, Kure S, Kirkner GJ, Schernhammer ES, Hazra A, Hunter DJ, Quackenbush J, Spiegelman D, Giovannucci EL, Fuchs CS, Ogino S: **Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample.** *PLoS One* 2008, **3:**e3698.
7. Brink M, de Goeij AF, Weijenberg MP, Roemen GM, Lentjes MH, Pachen MM, Smits KM, de Bruine AP, Goldbohm RA, van den Brandt PA: **K-ras oncogene mutations in sporadic colorectal cancer in The Netherlands Cohort Study.** *Carcinogenesis* 2003, **24:**703–710.
8. English DR, Young JP, Simpson JA, Jenkins MA, Southey MC, Walsh MD, Buchanan DD, Barker MA, Haydon AM, Royce SG, Roberts A, Parry S, Hopper JL, Jass JJ, Giles GG: **Ethnicity and risk for colorectal cancers showing somatic BRAF V600E mutation or CpG island methylator phenotype.** *Cancer Epidemiol Biomarkers Prev* 2008, **17:**1774–1780.
9. Fariña-Sarasqueta A, van Lijnschoten G, Moerland E, Creemers GJ, Lemmens VEPP, Rutten HJT, van den Brule AJC: **The BRAF V600E mutation is an independent prognostic factor for survival in stage II and stage III colon cancer patients.** *Ann Oncol* 2010, **21:**2396–2402.
10. Gavin PG, Colangelo LH, Fumagalli D, Tanaka N, Remillard MY, Yothers G, Kim C, Taniyama Y, Kim SI, Choi HJ, Blackmon NL, Lipchik C, Petrelli NJ, O'Connell MJ, Wolmark N, Paik S, Pogue-Geile KL: **Mutation Profiling and Microsatellite Instability in Stage II and III Colon Cancer: An Assessment**

of Their Prognostic and Oxaliplatin Predictive Value. *Clin Cancer Res* 2012, **18:**6531–6541.

11. Nakanishi R, Harada J, Tuul M, Zhao Y, Ando K, Saeki H, Oki E, Ohga T, Kitao H, Kakeji Y, Maehara Y: **Prognostic relevance of KRAS and BRAF mutations in Japanese patients with colorectal cancer.** *Int J Clin Oncol* 2012:1–7. http://dx.doi.org/10.1007/s10147-012-0501-x.
12. Ogino S, Shima K, Meyerhardt JA, McCleary NJ, Ng K, Hollis D, Saltz LB, Mayer RJ, Schaefer P, Whittom R, Hantel A, Benson AB, Spiegelman D, Goldberg RM, Bertagnolli MM, Fuchs CS: **Predictive and prognostic roles of BRAF mutation in stage III colon cancer: results from intergroup trial CALGB 89803.** *Clin Cancer Res* 2012, **18:**890–900.
13. Roth AD, Tejpar S, Delorenzi M, Yan P, Fiocca R, Klingbiel D, Dietrich D, Biesmans B, Bodoky G, Barone C, Aranda E, Nordlinger B, Cisar L, Labianca R, Cunningham D, Van Cutsem E, Bosman F: **Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial.** *J Clin Oncol* 2010, **28:**466–474.
14. Pai RK, Jayachandran P, Koong AC, Chang DT, Kwok S, Ma L, Arber DA, Balise RR, Tubbs RR, Shadrach B, Pai RK: **BRAF-mutated, microsatellite-stable adenocarcinoma of the proximal colon: an aggressive adenocarcinoma with poor survival, mucinous differentiation, and adverse morphologic features.** *Am J Surg Pathol* 2012, **36:**744–752.
15. Hutchins G, Southward K, Handley K, Magill L, Beaumont C, Stahlschmidt J, Richman S, Chambers P, Seymour M, Kerr D, Gray R, Quirke P: **Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer.** *J Clin Oncol* 2011, **29:**1261–1270.
16. Ren J, Li G, Ge J, Li X, Zhao Y: **Is K-ras gene mutation a prognostic factor for colorectal cancer: a systematic review and meta-analysis.** *Dis Colon Rectum* 2012, **55:**913–923.
17. Tejpar S, Bertagnolli MM, Bosman F, Lenz HJ, Garraway L, Waldman F, Warren R, Bild A, Collins-Brennan D, Hahn H, Harkin DP, Kennedy R, Ilyas M, Morreau H, Proutski V, Swanton C, Tomlinson I, Delorenzi M, Fiocca R, Van Cutsem E, Roth A: **Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery.** *Oncologist* 2010, **15:**390–404.
18. Van Cutsem E, Labianca R, Bodoky G, Barone C, Aranda E, Nordlinger B, Topham C, Tabernero J, Andre T, Sobrero AF, Mini E, Greil R, Di Costanzo F, Collette L, Cisar L, Zhang X, Khayat D, Bokemeyer C, Roth AD, Cunningham D: **Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3.** *J Clin Oncol* 2009, **27:**3117–3125.
19. Roth AD, Delorenzi M, Tejpar S, Yan P, Klingbiel D, Fiocca R, d'Ario G, Cisar L, Labianca R, Cunningham D, Nordlinger B, Bosman F, Van Cutsem E: **Integrated analysis of molecular and clinical prognostic factors in stage II/III colon cancer.** *J Natl Cancer Inst* 2012, **104:**1635–1646.
20. Tejpar S, Celik I, Schlichting M, Sartorius U, Bokemeyer C, Van Cutsem E: **Association of KRAS G13D tumor mutations with outcome in patients with metastatic colorectal cancer treated with first-line chemotherapy with or without cetuximab.** *J Clin Oncol* 2012, **30:**3570–3577.
21. Tejpar S, Popovici V, Delorenzi M, Budinska E, Estrella H, Mao M, Yan P, Weinrich S, Van Cutsem E, Roth A: **Mutant KRAS and BRAF gene expression profiles in colorectal cancer: Results of the translational study on the PETACC 3-EORTC 40993-SAKK 60-00 trial.** *J Clin Oncol* 2010, **28**(suppl):262s. abstr 3505. ASCO Annual Meeting abstract 3505.
22. Popovici V, Budinska E, Tejpar S, D'Ario G, Di Narzo AF, Hodgson JG, Roth A, Bosman F, Delorenzi M: **Molecular and clinicopathologic evidence of heterogeneity in KRAS-mutant colon cancers.** *J Clin Oncol* 2012, **30**(15_suppl):abstr 3575. ASCO Annual Meeting abstract 3575.

doi:10.1186/1471-2407-13-439
Cite this article as: Popovici *et al.*: Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer. *BMC Cancer* 2013 **13**:439.

[***9***] **Budinska E,** Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S, Bosman F, Roth A, Delorenzi M. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol. 2013 Sep;231(1):63-76. doi: 10.1002/path.4212. Epub 2013 Jul 8. PMID: 23836465; PMCID: PMC3840702.

Journal of Pathology
J Pathol 2013; **231:** 63–76
Published online 8 July 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.4212

ORIGINAL PAPER

Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer

Eva Budinska,[1,2]* Vlad Popovici,[1,2] Sabine Tejpar,[3] Giovanni D'Ario,[1] Nicolas Lapique,[1] Katarzyna Otylia Sikora,[1] Antonio Fabio Di Narzo,[1] Pu Yan,[4] John Graeme Hodgson,[5] Scott Weinrich,[5] Fred Bosman,[5] Arnaud Roth[6,7] and Mauro Delorenzi[1,8]

[1] *Bioinformatics Core Facility, Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland*
[2] *Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic*
[3] *Department of Oncology, University Hospital Gasthuisberg, Katholik Universiteit Leuven, Belgium*
[4] *University Institute of Pathology, Lausanne University Medical Centre, Switzerland*
[5] *Pfizer Inc., Worldwide Research and Development, Oncology Research Unit, La Jolla, CA, USA*
[6] *Oncosurgery, Geneva University Hospital, Switzerland*
[7] *Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland*
[8] *Département de Formation et Recherche, Lausanne University Medical Centre, Switzerland*

**Correspondence to: Eva Budinska, Institute of Biostatistics and Analyses, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic e-mail: budinska@iba.muni.cz*

Abstract

The recognition that colorectal cancer (CRC) is a heterogeneous disease in terms of clinical behaviour and response to therapy translates into an urgent need for robust molecular disease subclassifiers that can explain this heterogeneity beyond current parameters (MSI, *KRAS*, *BRAF*). Attempts to fill this gap are emerging. The Cancer Genome Atlas (TGCA) reported two main CRC groups, based on the incidence and spectrum of mutated genes, and another paper reported an EMT expression signature defined subgroup. We performed a prior free analysis of CRC heterogeneity on 1113 CRC gene expression profiles and confronted our findings to established molecular determinants and clinical, histopathological and survival data. Unsupervised clustering based on gene modules allowed us to distinguish at least five different gene expression CRC subtypes, which we call surface crypt-like, lower crypt-like, CIMP-H-like, mesenchymal and mixed. A gene set enrichment analysis combined with literature search of gene module members identified distinct biological motifs in different subtypes. The subtypes, which were not derived based on outcome, nonetheless showed differences in prognosis. Known gene copy number variations and mutations in key cancer-associated genes differed between subtypes, but the subtypes provided molecular information beyond that contained in these variables. Morphological features significantly differed between subtypes. The objective existence of the subtypes and their clinical and molecular characteristics were validated in an independent set of 720 CRC expression profiles. Our subtypes provide a novel perspective on the heterogeneity of CRC. The proposed subtypes should be further explored retrospectively on existing clinical trial datasets and, when sufficiently robust, be prospectively assessed for clinical relevance in terms of prognosis and treatment response predictive capacity. Original microarray data were uploaded to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under Accession Nos E-MTAB-990 and E-MTAB-1026.

Keywords: colorectal cancer; histopathology; gene expression; molecular heterogeneity

Received 3 February 2013; Revised 10 May 2013; Accepted 14 May 2013

Conflict of interest: ST and FTB have received research funds from Pfizer; and AR is a compensated consultant for Pfizer.

Introduction

Current classifications of sporadic colorectal cancer take into consideration stage, histological type and grade [1]. Colorectal cancer (CRC) is a highly heterogeneous disease, with clinicopathologically similar tumours differing strikingly in treatment response and patient survival. These differences are only partly explained by current concepts regarding the molecular events leading to CRC. In recent years, microsatellite instability (MSI) emerged as an important classifier with significant prognostic impact and potential for patient stratification for therapy [2,3]. Some molecular markers, as well as the mutation status of *BRAF* or *KRAS* genes (predictive for anti-EGFR [4]), are in use for treatment decisions and patient stratification. However, patient groups defined by these molecular markers still differ remarkably in behaviour and therapy response [5,6]. Several approaches to further subtype CRC have been proposed, based on combinations of clinical, histopathological, gene expression, CNV, epigenetic and single gene parameters [7–13]. Each of these different modalities provides its own perspective on the same underlying biological reality. The CpG island methylator phenotype (CIMP) status is emerging as important molecular determinant of CRC heterogeneity [11]. The cancer genome atlas (TCGA) analysis identified a hypermutant group not entirely captured by MSI status [13]. Several studies have addressed CRC subtyping using genome-wide gene expression profiling of relatively large patient cohorts [12,14]. One study used unsupervised clustering of stage II and III CRCs to identify three stage-independent subtypes, with *BRAF* mutation and MSI status dominating one of the subtypes [14]. A study of stage I–IV CRC samples segregated CRC into two prognostic subtypes with epithelial–mesenchymal transition (EMT) as a main determinant [12]. Another study on 88 stage I–IV samples identified four subtypes, one correlated with MSI, *BRAF* mutation and mucinous histology, two with stromal component and one with high nuclear β-catenin expression [15].

We recently reported CRC expressing a *BRAF*-mutated signature [6], which strongly overlaps with the methylation-based group of Hinoue [11], and a MSI-like gene expression group that captures the hypermutant tumours of TCGA [13], indicating the potential for identification of robust biological subgroups. We now describe CRC subtypes based upon unsupervised clustering of genome-wide expression patterns. We characterized these subtypes in terms of biological motifs, common clinical variables, association with known CRC molecular markers and morphological patterns. A key element in our approach was the use of a system of unsupervised gene modules—groups of genes with correlated expression. They are more resistant to noise and have a higher chance of having at least a few members represented on various platforms. In addition, as each gene module is represented by its median expression, the modules with fewer genes contribute equally to the subtype definition. We and others have successfully used similar strategies previously [16–18]. We validated the existence of the subtypes and their respective clinical and molecular marker characteristics in an independent dataset. Ultimately, it will be mandatory to integrate the various sources of information on CRC heterogeneity into an integrative, robust and reproducible subclassifier that can become a tool for clinical use.

Materials and methods

A detailed description of all the datasets and analysis procedures is given in Supplementary methods and results (see Supplementary material).

Data acquisition and processing

We have built two non-overlapping data collections: a discovery collection, comprising four publicly available (425 samples) and two previously unpublished datasets (688 samples with 10 year follow-up in a clinical trial setting and 64 normal samples) with known stage status, and a validation collection of eight publicly available datasets (720 CRC samples) (see Supplementary material, Supplementary methods and results). Observations derived from the analysis of 64 normal samples were further validated on five publicly available datasets, with both carcinoma and normal samples available in one batch (totalling 205 normal/adenoma/carcinoma samples). Copy number data was available for 154 of the PETACC3, as in [19]. Our analysis included a total of 2102 samples.

The discovery collection contained the previously unpublished 688 CRC formalin-fixed, paraffin-embedded (FFPE) samples of PETACC3 [6] and 64 FFPE normal colon tissue samples from Centre Hospitalier Universitaire Vaudois's Biobank, which were uploaded to ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), under Accession Nos E-MTAB-990 and E-MTAB-1026, respectively. Gene expression data were processed by standard tools to obtain normalized, probeset-level expression data. For each EntrezID in the datasets, the probeset with the highest variability was selected as representative and the number of EntrezIDs entering the analysis was reduced to 3025 by applying non-specific filtering. For PETACC3 and normal colon samples, patients signed an informed consent form in which the use of tissue specimens was included, and all marker study proposals were subjected to the approval of the trial steering committee.

Subtype definition and validation

For model development (gene modules and subtype definition, classifier training, identification of subtype-specific genes) only the 1113 CRC samples of the discovery set were used, no sample in the validation collection being used for any model tuning. Hierarchical clustering (complete linkage, Pearson correlation similarity measure) and dynamic cut tree [20] were used to produce *gene modules* (groups of genes with correlated expression), from which non-robust modules (see Supplementary material, Supplementary methods and results) and a gender-related module were discarded. Each expression profile was then reduced to a vector of *meta-gene*s by taking the median of the values of genes in each gene module. The meta-genes were then further grouped into clusters using hierarchical clustering.

The subtypes were defined in terms of *core samples*—those samples from the discovery collection that were assigned to clusters by hierarchical clustering, using a consensus distance [21] followed by pruning of the dendrogram (see Supplementary material, Supplementary methods and results). The clusters to which the core samples were assigned were called *subtypes*. The rest of the samples from the discovery collection, not assigned to subtypes by this procedure, were called *non-core samples*. This approach allowed the reduction of noise in subtype-defining samples, and thus a higher consistency of the resulting subtypes defining the ground truth for downstream analyses. The stability of the obtained clusters was assessed under different perturbations of the processing pipeline (different parameters and clustering methods) to ensure that the results were not simple artefacts (see Supplementary material, Supplementary methods and results). A multiclass linear discriminant (LDA) [22] was trained on core samples with meta-genes as variables to assign new samples to one of the subtypes. Minimal gene sets characteristic to each subtype were identified using ElasticNet [23] on gene-level data.

In order to validate the existence of subtypes (and their independence on data selection) and the modelling choices in subtype discovery, we applied the same subtyping procedure (including parameters) to the validation collection. The clusters identified in the validation collection were put in correspondence with the subtypes in the training set by LDA predictions and correlations of subtype-specific moderated *t* statistic [24] values, corresponding to the gene-wise comparison of the respective subtype with the other subtypes (one-versus-all comparison). A simple classifier application would have led the validation samples to be classified as one of the subtypes, but it would have not informed us of possible over-fitting of the data in the discovery procedure.

Subtype characterization

If not specified differently, all the reported *p* values were adjusted for multiple hypothesis testing, using the Benjamini–Hochberg procedure. Significance level was set at 0.1. Pathway analysis for each set of gene modules was carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [25]. Gene set enrichment analysis of gene signatures was performed using the mygsea2 tool, in each subtype and normal samples, on average expression-ordered median-centred lists of genes. Differential expression analysis was performed using limma [24] and sign test using BSDA [26]. The Cox proportional hazards model was used to analyse the prognostic value of interquartile range (IQR)-standardized values of meta-genes, for overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR), stratified by dataset. The Wald test was used to assess the global significance of the models. Pairwise differences in survival were assessed using the log-rank test. For subtype comparison, the survival was truncated at 7 years. Subtype enrichment for clinical or molecular markers was assessed by the Fisher test to the baseline, defined as the proportion of the marker in the whole dataset. Morphological pattern differences were assessed pairwise by Fisher test.

Histology

The identified subtypes were characterized histologically in terms of six different architectural patterns: complex tubular; solid/trabecular; mucinous; papillary; desmoplastic; and serrated (Figure 4A), which were called dominant or secondary depending on their presence in the histology slides (for details on immunohistochemistry, see Supplementary material, Supplementary methods and results).

Results

Gene modules and subtype definition

We identified 54 gene modules, reproducible across all datasets in the discovery collection, comprising 658 genes from an initial list of 3025 identified as the most variable. The assignment of genes to gene modules and gene module clusters is listed in Table S1 (see Supplementary material); meta-gene expression profiles for the discovery set are shown in Figure 1A; and between meta-gene correlations in Figure S1C (see Supplementary material). Based on gene modules, we identified five major subtypes: surface crypt-like (A), lower crypt-like (B), CIMP-H-like (C), mesenchymal (D) and mixed (E), totalling 765 samples (69% of discovery data; see Supplementary material, Supplementary methods and results).

Subtype reproducibility in an independent validation set

In the validation set of 720 CRC samples we identified a set of subtypes comprising 602 samples (83.6% of the validation set) and associated them with our discovery subtypes using the subtype classifier (see Supplementary material, Table S2) and correlations of subtype-specific patterns based on moderated *t* statistic (see Supplementary material, Table S3). All five major subtypes reappeared in the validation set, confirming the robustness of our approach. Figure S2 (see Supplementary material) presents gene expression profiles of both discovery and validation sets. Two notable differences were observed: (i) subtype B in the validation set was split into two subgroups (B1, B2), as observed in the discovery set too, but only at lower pruning height; (ii) another cluster passed the minimal size criteria, corresponding to the small subtype (F) which, in the discovery set, was not considered for further characterization because of small sample size. Validation of other subtype characteristics (to the extent of available information) is described in each of the respective sections.

Subtypes are characterized by distinct biological components

We set out to assign biological labels to gene modules that define the subtypes (Table 1; see also Supplementary material, Table S1). Of the 54 meta-genes,

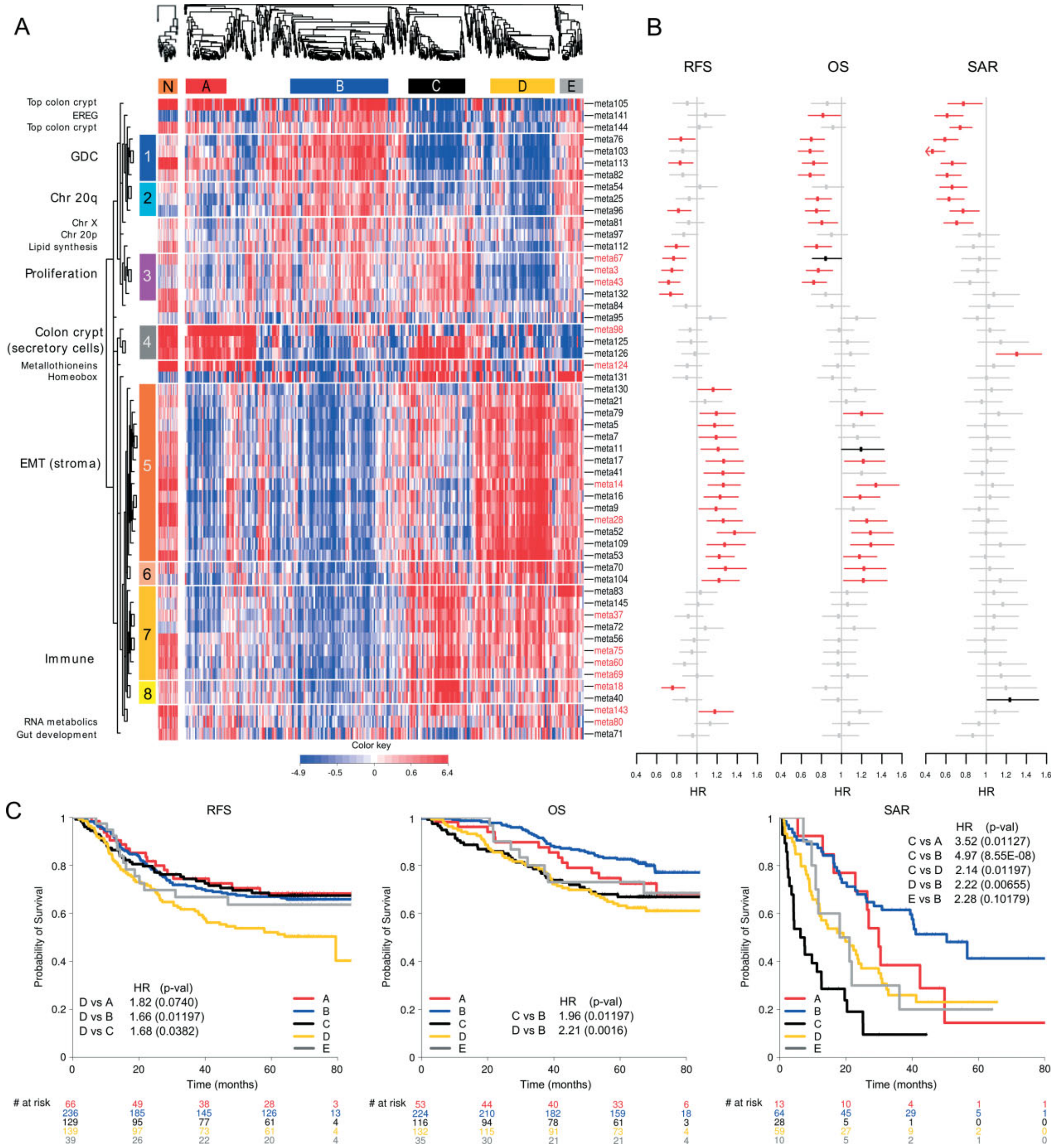


Figure 1. Meta-gene expression pattern in subtypes, connected with prognostic effect of subtypes and meta-genes, in the discovery set. (A) Two heat maps clustering normal (left) and CRC (right) samples (columns) and meta-genes (rows). Colours represent decreased (blue) or increased (red) meta-gene expression relative to their medians. Normal samples were clustered independently on meta-genes centred to CRC meta-gene medians. For comparative purposes, ordering of meta-genes in normal samples is imposed to correspond to that of CRC samples. White horizontal lines denote eight unsupervised clusters of meta-genes, each assigned a colour bar on the left; meta-genes not belonging to a cluster have no colour bar. Names of the meta-genes corresponding to gene modules with gene–gene correlations in normal samples comparable to those in cancer samples are marked red (see Supplementary material, Figure S1D). (B) Effect of inter-quartile range (IQR) standardized expression of meta-genes on RFS, OS and SAR. Points represent estimated hazard ratio (HR), bars represent 95% CI. Bold lines represent effects significant at 5% without adjustment for multiple hypothesis testing; red lines represent effects significant at FDR < 10%; details are provided in Table S6 (see Supplementary material). (C) Kaplan–Meier plots for RFS, OS and SAR, with HR for significant pairwise comparisons (*p* values adjusted for FDR). Numbers below *x* axes represent number of patients at risk at selected time points.

41 could be further grouped into eight gene module clusters; 13 meta-genes remained ungrouped, each possibly representing a distinct biological motif. Pathway analysis characterized five of eight gene module clusters by the following biological motifs: chromosome 20q (cluster 2), proliferation (cluster 3), EMT/stroma (cluster 5) and immune response (clusters 7 and 8). Literature searching identified biological motifs associated with other gene modules. We labelled cluster 1 as GDC (genes differentially expressed in CRC), as

Table 1. Biological identification of gene modules

Cluster name	Number of genes	Pathway analysis result (number of overlapping genes, *p* value) OR description based on literature search	Selected genes
1. GDC	27	Genes involved in differentiation of colon crypt and/or whose expression was reported to be affected in colorectal cancer and/or with prognostic effect in CRC	Intestinal differentiation genes: *CDX2*[45], *IHH*[46], *VAV3*[47], *ASCL2*[35], *PLAGL2*[48] Genes reported altered in colorectal cancer with prognostic effect: *PITX2*[49], *DDC*[50], *PRLR*[51], *SPINK1*[52] Other genes connected to CRC: *GGH*–connected to CIMP^{+} phenotype [53] *NR1I2*–connected to chemoresistance [54]
2. Chromosome 20q genes	33	Chromosome 20 (26 genes, 9.2E-34)	Other, non-20q genes: *TP53RK*, *ANO9*, *NEU1*, *CLDN3*, *PRSS8*
3. Proliferation	83	Cell cycle (36 genes, 3.0E-33) Mitosis (26 genes, 1.4E-29) Chromosome (26 genes, 2.5E-17) DNA metabolic process (20 genes, 4.9E-10) Lipid synthesis (4 genes, 5.0E-2)	Mitotic checkpoint kinases: *BUB1*, *BUB1B* Cyclins: *CCNA2*, *CCNB2* Centromere proteins: *CENPA*, *CENPE*, *CENPN* Kinesins: *KIF11*, *KIF23*, *KIF4A* Topoisomerase II (*TOP2A*) Cell division cycle 2 *CDC2*
4. Colon crypt markers (secretory cells)	16		*AGR2*[55], *AGR3*, *MUC2*, *SPINK4*[56], *RETNLB*[57], *REG4*[58]
5. EMT/stroma	310	Extracellular region part (90 genes) 2.7E-36 Cell adhesion (57 genes) 1.2E-17 Extracellular matrix (44 genes) 5.3E-30 Collagen (16 genes) 1.2E-15 EGF-like domain (26 genes) 1.6E-12 Cell motion (33 genes) 7.2E-8 Blood vessel development (25 genes) 1.1E-8 Growth factor binding (6 genes) 6.0E-5 Frizzled related (5 genes) 6.7E-3 Cell junction organization (7 genes) 1.8E-2 WNT receptor signalling pathway (8 genes) 1.4E-1	Inhibitors of β-catenin-dependent canonical WNT: *SFRP1*, *SFRP2*, *SFRP4*, *DKK3*, *FZD1,7*, *PRICKLE1*, *NXN* Mesenchymal markers: N-cadherin, OB cadherin, *SPARC*, *DDR2* EMT inducers(TFs): *SNAI2*, *ZEB1*, *ZEB2*, *TWIST1*, *CDH11* ECM remodelling and invasion: *MMP14*, *VIM* ECM proteins: fibronectin 1, collagens Angiogenesis: *PLAT*, *PLAU*, *NRP1*, *NRP2*, *THBS1*, *THBS2*, *THBS4* TGFs, their receptors and binding proteins: *IGF1*, *IGFBP5*, *IGFBP7*,*TGFB*, *LTBP1*, *LTBP2*, *PDGFRA*, *PDGFRB*
6. Unidentified	14		*DUSP1*, *EGR2*, *SERPINE1*
7 and 8. Immune response	103	Immune response (42 genes) 2.0E-28 Positive regulation of immune system process (16 genes) 4.0E-9 Antigen processing and presentation via MHC class II (6 genes) 7.5E-5 Defence response (31 genes) 3.3E-17 Chemokine signalling pathway (9 genes) 2.2E-3 Lymphocyte activation (11 genes) 2.1E-5 Regulation of programmed cell death (14 genes) 2.1E-2	Cytokines: *CCL3*, *CXCL5*, *CXCL9*,*CXCL10*, *CXCL11*, *SPP1*, *LTB* MHC class II: *HLA-DMB*, *HLA-DPA1*, *HLA-DRA*, *CD74* MHC class I: *HLA-F*, *TAP1*, *TAP2* Anti-apoptotic: *BCL2A1*, *CD74*, *BIRC3*, *IFI6*, *TNFAIP3*, *TNFAIP3* Apoptotic: *STAT1*, *XAF1* Interferon-induced proteins: *IFI30*, *IFI16*, *IFI44*, *IFI16*, *IFIH1*, *IFIT3*
Cluster-unassigned meta-genes with colon crypt cell markers (enterocytes/top of the crypt)			
Meta-gene 105	6	Top of the crypt genes	*FAM55A*, *FAM55D*, *MUC12* and *CEACAM7*[59], *SLC26A2*[59], *SLC26A3*[59]
Meta-gene 144	5	Enterocytes, goblet cells markers	*LOC644844*, *NGEF*, *HEPH*, *KRT20*[59], *MUC20*[59]
Cluster-unassigned meta-genes associated with chromosomal location 0			
Meta-gene 81	7	Chromosome X (7 genes) 1.1E-8	*CXorf15*, *EIF1AX*, *HDHD1A*, *MED14*, *PNPLA4*, *SCML1*, *SMC1A*
Meta-gene 97	6	Chromosome 20p (5 genes) 5.0E-11	*CDC25B*, *CSNK2A1*, *MRPS26*, *PTPRA*, *RP5-1022P6.2*, *SNRPB*
Meta-gene 84	7	Chromosome 8 (7 genes) 5.4E-9	*AGPAT5*, *FDFT1*, *GTF2E2*, *LONRF1*, *MTUS1*, *VPS37A*, *ZNF395*
Other cluster-unassigned meta-genes			
Meta-gene 141	5	EREG	*AK3L1*, *ARID3A*, *EREG*, *LDLRAD3*, *ZBTB10*
Meta-gene 112	6	Lipid synthesis (4 genes) 5.0E-2	*DHCR7*, *FASN*, *FGFBP1*, *HMGCS1*, *IDI1*, *PCSK9*
Meta-gene 95	6	Homeobox genes	*HOXA10*, *HOXA11*, *HOXA13*, *HOXA5*, *HOXA7*, *HOXA9*
Meta-gene 124	5	Metallothioneins	*MT1E*, *MT1F*, *MT1G*, *MT1M*, *MT1X*
Meta-gene 131	5	Disulphide bonds (5 genes) 1.7E-02	*CXCL5*, *IL6*, *MMP1*, *MMP3*, *PTGS2*
Meta-gene 143	5	Unidentified	*DUSP5*, *ERRFI1*, *KLF6*, *MXD1*, *PLAUR*
Meta-gene 80	7	Regulation of RNA metabolic process (6 genes) 4.9E-2	*ATF3*, *C8orf4*, *FOS*, *JUNB*, *NR4A1*, *SIK1*, *ZFP36*
Meta-gene 71	8	Gut development (3 genes) 3.5E-2	*CCL11*, *CH25H*, *EDNRB*, *F2RL2*, *FOXF1*, *FOXF2*, *PCDH18*, *WNT5A*

Table 2. Subtype-specific minimal gene set as identified by Elastic net

	Minimal gene sets specifying a subtype	
Subtype	Up-regulated from population mean	Down-regulated from population mean
A. Surface crypt-like	*ADTRP*, *B3GNT7*, *CLCA1*, *MUC2*, *NR3C2*, *PADI2*, *RETNLB*, *STYK1*	*CHI3L1*, *FNDC1*, *TIMP3*, *SULF1*
B. Lower crypt-like	*CCDC113*, *CDHR1*, *FARP1*, *GPSM2*, *GRM8*, *HNF4A*, *IHH*, *KCNK5*, *KIAA0226L*, *MYRIP*, *PLAGL2*, *PRR15*, *QPRT*, *RNF43*, *RPS6KA3*, *SLC5A6*, *TP53RK*, *TSPAN6*, *VAV3*, *YAE1D1*	*ALOX5*, *BASP1*, *CREB3L1*, *CXCR4*, *EPB41L3*, *FSCN1*, *GFPT2*, *GPX8*, *ITPRIP*, *KCNMA1*, *KCTD12*,*MT1E*, *RARRES3*, *RNASE1*, *SGK1*, *SOCS3*
C. CIMP-H-like	*ANP32E*, *EGLN3*, *IDO1*, *PLK2*, *RAB27B*, *RARRES3*, *RPL22L1*, *TFAP2A*	*ATP9A*, *C10orf99*, *CXCL14*, *KIAA0226L*
D. Mesenchymal	*ANK2*, *BOC*, *C7*, *CRYAB*, *DCHS1*, *DDR2*, *GEM*, *PRICKLE1*, *TAGLN*	*HOOK1*, *RBM47*
E. Mixed	*CEACAM6*, *CXCL5*, *HSD11B1*, *IL1B*, *IL6*, *MRPS31*, *PI15*, *RAP2A*, *UQCC*	*AGR3*, *RAB27B*, *REG4*

it consisted of a number of genes significantly associated with CRC. The analysis of pairwise intra-gene module correlations in normal samples of both discovery and validation set identified as cancer-specific gene modules of chromosome 20q, several immune response, EMT/stroma and GDC gene modules, homeobox genes and gut development (see Supplementary material, Figure S1D). The relationship between subtypes and meta-genes is illustrated by the heat map (Figure 1A), in which the major molecular motifs and their role in subtype definition stand out. Table S4 (see Supplementary material) contains median subtype values per meta-gene and the results of differential meta-gene expression testing between subtypes. Subtypes are not determined by individual biological components but each of them contributes to the molecular identity of the subtypes. The EMT/stroma cluster stands out in subtypes A + B (low expression) and D + E (high expression), while subtype C notably contained a high expression of immunity-associated cluster. High expression of meta-genes representing upper colon crypt cells in subtypes A and B, correlated with serrated and papillary (A) and complex tubular (B) morphological patterns (see below). Given the enterocyte-like morphology and retained polarity of the neoplastic cells in these patterns, they are considered as well differentiated. Subtype C is associated with the mucinous phenotype. Interestingly, subtypes A and C show high expression of metallothioneins, subtypes C and E show high expression of the homeobox gene module, while subtypes E and B strongly express a gene module containing the *EREG* gene (Table 1). The high expression of chromosome 20q cluster in subtype B was correlated with a significantly higher copy number gain/amplification of all of 20q in this subtype (see Supplementary material, Figure S8). The low expression of lipid synthesis genes is striking for subtype D and low expression of the gut development gene module for subtype C. A refined picture of differences is given by a quantitative comparison of (meta-)gene expression between subtype pairs (see Supplementary material, Tables S4 and S5, Figure S4). For each subtype we also identified a minimum set of characteristic genes (Table 2; for more details, see Supplementary material, Supplementary methods and results).

Normal colon mucosa in the context of subtypes

When applied to the 64 normal samples, the LDA classifier assigned them all to subtype A, with posterior probability > 0.99, supporting the observation that A is well differentiated and closest to normal colonic epithelium in terms of gene expression pattern. For validation, we analysed five public datasets comprising 205 profiles of normal/adenoma/carcinoma samples. Most of the normal and adenoma samples were classified by LDA as subtype A (74.5% of 51 and 69.0% of 71, respectively) or subtype B (28.2% and 21.6%, respectively), confirming subtype A as the most normal-like. The 80 carcinoma samples were distributed over all subtypes (26.2% A, 30.0% B, 11.3% C, 18.7% D and 13.8% E).

Subtypes and patient survival

We assessed whether subtypes differ in survival, as a general read-out of biological significance, and then tested the association of each meta-gene with prognosis, using the complete discovery set of 1113 patients (Figure 1B-C see also Supplementary material, Table S6). Kaplan–Meier curves for RFS, OS, SAR, hazard ratios (HRs) and p values of pairwise differences between subtypes are shown in Figure 1C. The results indicate that subtypes C and D are associated with poor OS. For subtype D, this is primarily due to early relapse correlated with high expression of EMT genes and low expression of proliferation-associated genes. For subtype C it is the result of short SAR, correlated with low expression of GDC, top colon crypt, EREG and Chr 20q genes and high expression of meta-gene 126 (see Supplementary material, Table S1). For subtype E the trend towards poorer OS and RFS was not statistically significant, although borderline significant poorer SAR was found relative to subtype B. Subtypes A and B had better prognosis than D for all three endpoints, although for OS in subtype A this was not significant.

The analysis of clinical and molecular markers (below) showed that subtype C is enriched for MSI tumours and *BRAF* mutant tumours, the latter present also in subtype D. The literature indicates that MSI is associated with better RFS, while *BRAF* mutation is an indicator of worse SAR [27]. To analyse how these two contradictory components affect survival in

Table 3. Result of additive multivariate Cox proportional hazards model, with subtype, *BRAF* mutation, MSI and stage[a]

Variable	RFS HR	*p*	OS HR	*p*	SAR HR	*p*
A	0.906	0.760	1.381	0.390	1.726	0.180
C	0.940	0.850	1.560	0.220	3.675	0.0022*
D	1.688	0.0055*	2.161	0.0011*	1.906	0.014*
E	1.506	0.210	2.201	0.035*	2.046	0.075
BRAFm	1.633	0.085	2.472	0.0034*	3.361	0.00072*
MSI	0.478	0.044*	0.275	0.004*	0.356	0.036*
Stage 3	0.770	0.190	0.943	0.820	1.780	0.062*

[a] Baseline is subtype B, MSS, *BRAF* wt and Stage 2.

*Variables significant in the model.

Hazard ratios (HR) for relapse-free survival (RFS), overall survival (OS) and survival after relapse (SAR).

subtypes, we built a multivariate Cox proportional hazard model with subtype, stage, *BRAF* and MSI (Table 3; see also Supplementary material, Table S6). Subtype C remained significantly associated with poor SAR, even after the adjustment for *BRAF*, MSI and stage, but not with RFS. Subtypes B and D remained significantly prognostic for RFS, OS and SAR. No equivalent survival data were available for the datasets in the validation series, hence these observations could not be validated.

Colorectal stem cell and Wnt signatures within subtypes

We investigated the association of subtypes with Wnt [28–32], putative colon cancer stem cell (CSC) [33–35] signatures, and two signatures specific for upper and lower colon crypt compartments [36], using gene set enrichment analysis (Figure 2; see also Supplementary material, Table S7). Subtypes B and E highly expressed canonical Wnt signalling target signatures. Subtypes A and D and also normal samples, however, showed low expression of these signatures. This was in concordance with the differences in β-catenin nuclear immunoreactivity at the invasion front (IF; see Supplementary material, Figure S9 and Supplementary methods and results). Subtypes B and E showed the highest percentages, while subtypes A and D showed significantly lower percentages of the β-catenin-positive nuclei. Subtype C exhibited almost no β-catenin nuclear immunoreactivity at the IF. We analysed CSC signatures derived from low colon crypt compartment cells that had been identified either by a Wnt reporter construct TOP GFP or by high surface expression of *EphB2*. Subtypes D and E expressed both TOP GFP and *EphB2*-derived CSC signatures, while subtype B mainly expressed only the TOP GFP signature (Figure 2).

Subtypes complement clinical and molecular markers

An important goal of this study was to assess how our molecular subtypes complement known clinical variables and molecular markers. We found that MSI, *BRAF* mutation status, site, mucinous histology and expression of p53 were significantly associated with various subtypes (Figure 3), but not tumour stage, age, gender, *SMAD4* or *PIK3CA* mutations (see Supplementary material, Figure S5A). Subtype D was not significantly enriched for any of the tested variables except for the *BRAF* mutated signature and possibly represents a mixture of tumours that have the EMT/stroma signature in common. *KRAS* mutants occurred in all subtypes (see Supplementary material, Figure S5C), supporting the emerging notion that *KRAS*-mutated CRC are substantially heterogeneous [5,6,37], the oncogenic role of *KRAS* varying per specific mutation and the molecular background of the tumour in which it occurs [38]. Subtype C expressed the *BRAF* mutant signature we identified earlier [6] (87.0%), a CIMP-H signature ([11], Figure 2), and its characteristics (enrichment for MSI, right side and mucinous histology) corresponded with those of the previously reported CIMP-H phenotype [9,11,39,40] and hypermutated tumours [13]. Regarding the latter, subtype C had a similar low frequency of copy number variations (see Supplementary material, Figure S7). The distribution of MSI status, stage, age, gender, grade and site over the subtypes in the validation set followed the same patterns established in the discovery set [cf Figures 3 and S5B (see Supplementary material)]. A classification tree, trained with a combination of available clinical and molecular markers, did not identify our subtypes (see Supplementary material, Figure S5D), indicating that gene expression patterns reveal a layer of heterogeneity that goes beyond conventional CRC classification approaches.

Histological characteristics of subtypes

To study whether or not our molecular subtypes are associated with histological patterns, we examined haematoxylin and eosin (H&E)-stained paraffin sections of a randomly selected subset of each subtype (23, 31, 31, 29 and 19 cases for subtypes A, B, C, D and E, respectively). In attempting to match histological morphotypes to molecular subtypes, architectural patterns were used, as illustrated in Figure 4A, rather than the recognized WHO classification of CRCs [1]. Not surprisingly, given intratumour heterogeneity, none of the tumours had a single pattern. However, the prevalent patterns showed appreciable differences between the subgroups (Figure 4B, C; see also Supplementary material, Figure S6). In subtype A, the serrated pattern was most frequent, followed by the papillary pattern; in

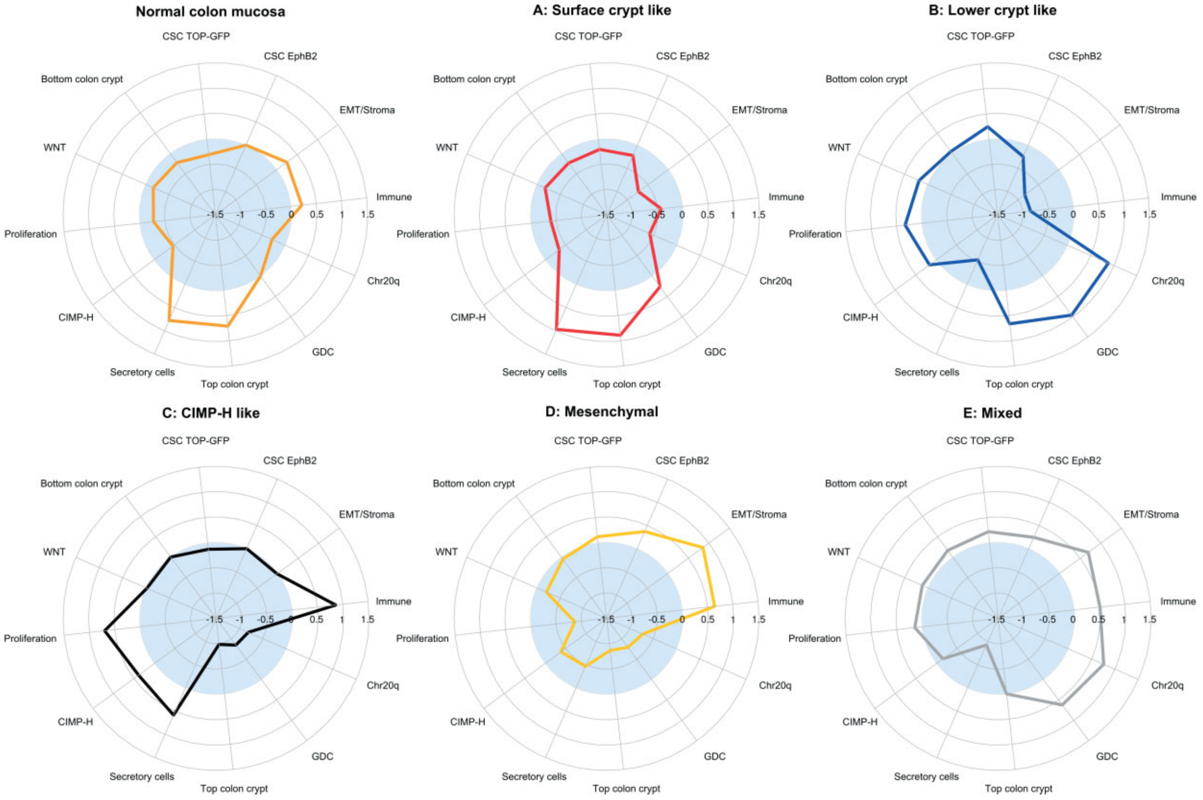


Figure 2. Subtypes and biological motifs. Subtype-specific fingerprints of biological motifs, represented either as mean values of gene set enrichment scores of gene sets from corresponding gene modules (EMT/stroma, immune, secretory cells, proliferation, GDC, chromosome 20q, top of the crypt—meta105 and meta144) or composed gene set enrichment scores of particular signatures (canonical Wnt targets, CSC-TopGFP, CSC-EphB2, colon crypt bottom and CIMP-H). The gene set enrichment scores represent whether the genes from the gene set show statistically significant enrichment between the down-regulated (negative scores, light blue area) or up regulated (positive scores) genes of a given subtype; details of score calculation can be found in the Supplementary material (Supplementary methods and results and Table S7.).

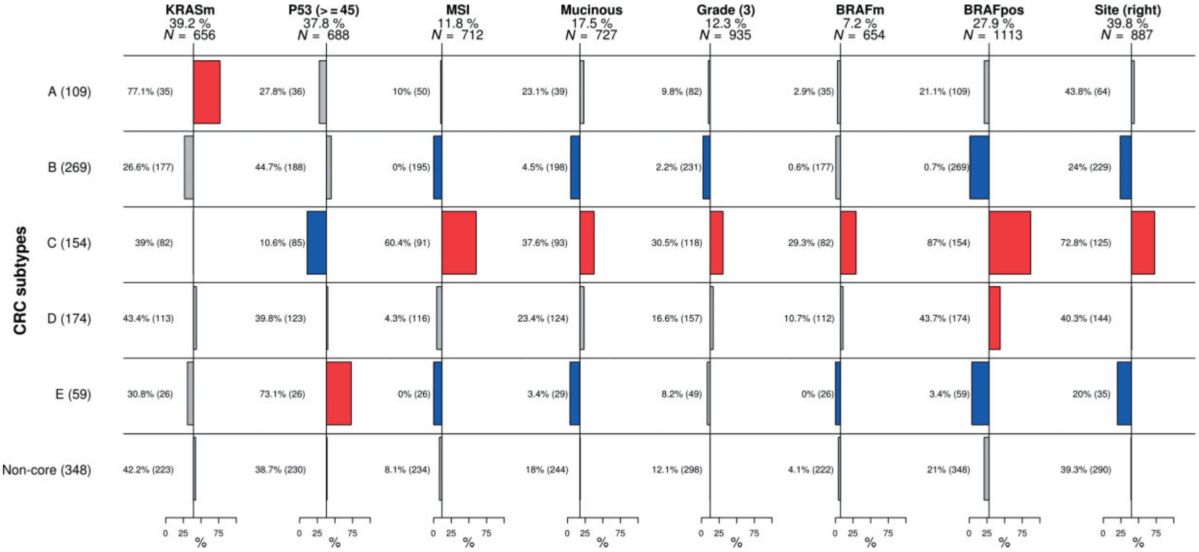


Figure 3. Clinical and mutational characterization of subtypes. Columns represent variables and rows subtypes. Horizontal bar plots represent proportions of the corresponding variable in each of the subtypes and non-core samples. Non-core samples were tested as one group to ensure that they did not share a common characteristic that would set them apart. Numbers in brackets adjacent to subtype name represent overall number of samples in the subtype. Under the title of each variable we denote the percentage representing baseline proportion in the population, with available information, and *N* denotes the number of patients for which the information on the respective feature was available. Bars in red represent significant enrichment and bars in blue significant depletion of a feature in the subtype in comparison to baseline, at the 5% significance level. Adjacent to each bar is the percentage of samples in the subtype with the specific feature and in brackets the overall number of samples in the subtype with the information available. We can read that, for instance, subtype C, comprising 154 samples, is enriched for microsatellite-unstable (MSI) tumours, where 60.4% of 91 samples with available information are MSI.

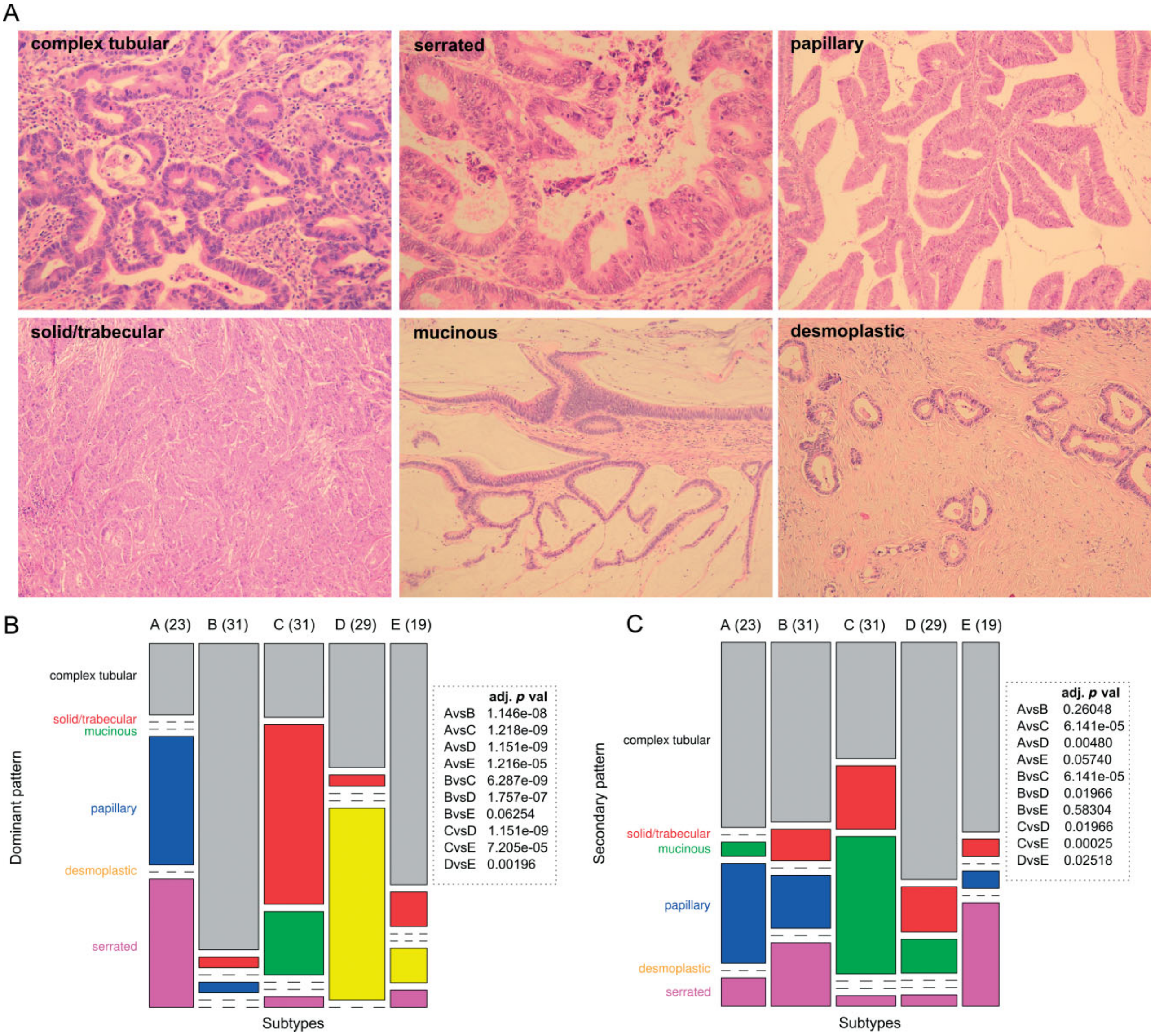


Figure 4. Morphological CRC patterns. (A) morphological CRC patterns scored in subtypes. (B, C) Distribution of dominant (B) and secondary (C) histological patterns in subtypes. Columns represent subtypes and widths are proportional to subtype frequency (numbers of samples in each subtype); rows represent dominant (B) or secondary (C) patterns and heights are proportional to pattern frequency. Boxes show adjusted *p* values of pairwise statistical testing of morphological pattern distribution between subtypes.

subtypes B and E, complex tubular dominated; in subtype C the solid pattern dominated, with mucinous as the second; most striking was the presence of a strong stromal reaction in subtype D.

Discussion

Our approach, using gene modules on a large panel of samples, allowed us to identify five main CRC gene expression subtypes (Table 4). It is relevant to note that subtyping can be performed on FFPE tissues, an important prerequisite for wide clinical applications. An example is the hypermutated group identified in the TCGA study by whole exome sequencing [13], but according to our data also by gene expression profiling on routinely processed tissues (CIMP-H-like subtype).

The combination of gene expression, clinical, mutational, survival and morphological data contributes new insight into the heterogeneity of CRC. While the validation confirmed the robustness of our findings across different platforms (ALMAC versus Affymetrix), sample preparation methods (FFPE versus fresh-frozen) and dataset collections, larger datasets are necessary to assess and characterize the relevance of lower frequency subtypes (eg F, or further segregation of B into B1 and B2). Our data indicate that several major biological processes are key determinants of a complex subtype structure of CRC. Therefore our subtypes defined by gene expression do not substitute but complement groups defined by current clinicopathological variables and molecular markers. Notably, morphological subclassification of CRC has clearly reached its limits, given the often striking intratumour

Table 4. Summary of subtype characteristics

	CRC markers and mutations				Histopathology	IHC	Median survival (months)			Clinical		Gene expression	
Subtype	MSI	*BRAF*	*KRAS*	*P53*	Dominant	Nuclear β-catenin at IF	OS	RFS	SAR	Site	Grade	Up-regulated	Down-regulated
A: Surface crypt-like		–	+		Papillary or serrated	–	NA	NA	28.9			Top colon crypt, secretory cell, metallothioneins	EMT/stroma, Wnt, CSC, Chr20q, proliferation
B: Lower crypt-like	–	–			Complex tubular	+	NA	NA	50.4	Left	2	Top colon crypt, proliferation, Wnt	EMT/stroma, immune, secretory cell
C: CIMP-H-like	+	+		–	Solid/trabecular or mucinous	–	NA	NA	6.9	Right	3	Proliferation, immune, metallothioneins	GDC, top colon crypt, Chr20q
D: Mesenchymal					Desmoplastic	–	NA	79.5	19.8			EMT/stroma, CSC, immune	Proliferation, secretory cell, top colon crypt, GDC, Wnt, Chr20q
E: Mixed	–	–		+	Complex tubular	+	NA	NA	19.6	Left		EMT/stroma, immune, top colon crypt, Chr20q, GDC, CSC	Secretory cell

+, significantly enriched; –, significantly depleted; IF, invasion front; NA, not attained; no value, no significant enrichment in comparison to population baseline.

heterogeneity, which made us use a (primary and secondary) architectural pattern approach rather than the canonized histological subtypes (WHO). Profiling of microdissected patterns within a single tumour might reveal molecular mechanisms responsible for these morphotypes. This additional heterogeneity within the subtypes may reflect tumour polyclonality, similar to breast cancer [41]. Ultimately, aggregating clinical, pathological and further detailed molecular characteristics (including CNV, miRNA and methylation) will contribute to a more detailed perception of CRC heterogeneity and it is likely that more subtypes will emerge. This, however, would need more detailed molecular annotation of larger clinically well documented CRCs.

A striking association was found between the stromal subtype D and the EMT signature. The previously discovered EMT [12] also emerged from our analysis as the largest cluster of meta-genes associated with poor RFS (subtype D). Our histological assessment suggests that the EMT signature is the reflection of a strong mesenchymal stromal reaction, and this histological characteristic deserves to be tested for its capacity to predict resistance to therapy, in view of its strong association with poor survival. Studies requiring high tumour cell content as sample inclusion criteria (eg [13]) could miss this poor prognosis subtype. Identification of this subtype in cell lines or xenograft models is less straightforward and would benefit from the analysis of gene expression patterns between microdissected tumour and stromal cells.

EMT, however important, only partly explains CRC heterogeneity, as even subtypes with similar expression of EMT-associated genes (A–C or D–E) differ in survival, mutational, clinical and gene expression characteristics. Additional biological components, such as differentiation, immune response, proliferation, chromosome 20q or cluster of genes deregulated in CRCs, are important co-determinants that underpin a need for further subdivision of CRCs. The findings from the analysis of CSC and WNT signatures support the recently suggested hypothesis that the colon stem cell signature under the condition of silenced canonical WNT targets is associated with higher risk of recurrence (subtype D) [33]. This is consistent with subtype D showing a significantly lower percentage of β-catenin-positive nuclei than subtype B, with its Wnt-associated gene expression and better survival.

MSI tumours represent a subclass in most unsupervised analyses and can be recognized at the gene expression level [42]. The more recent gene expression studies [14,15] suggest that MSI and *BRAF* share distinct gene expression patterns. Subtype C was enriched for both MSI and *BRAF* mutants and had one of the best outcomes for RFS, but the worse outcome in SAR, in concordance with previously reported results [43]. Subtype C retained its poor SAR prognostic value, even in the population of MSS and *BRAF* wild-type patients. Our data suggest that subtype C represents tumours with a common biology and a gene expression pattern that might best characterize a group of tumours resistant to chemotherapy, once metastatic. In this sense, our work not only agrees with the current known markers (*BRAF* mutation status and MSI) but clearly adds new insight, putting together these previously unrelated clusters into one biologically meaningful group. This observation is in line with recently published work [6].

Our observations show that gene expression profiling contributes substantially to our insight into CRC heterogeneity in confirming and complementing data from sequencing, CNV and promoter methylation analysis. Our subtypes can be further functionally interrogated for driving oncogenes/events by *in vitro* functional screens. High-risk subtypes D and C might contribute to therapeutic decision making in either adjuvant or metastatic settings. Retrospective analysis of clinical trial series may identify drug sensitivity associated with particular subtypes, and might open new treatment optimization strategies to be tested in clinical trials with stratified cohorts, similar to the I-SPY2 trial for breast cancer [44].

In conclusion, our unsupervised approach using gene modules resulted in the identification of distinct molecularly defined CRC subtypes, which adds a new layer of complexity to CRC heterogeneity and opens new opportunities for understanding the disease. The challenge is now to assimilate conventional and these new molecular approaches into a comprehensive consensus classification, which might then be used in further clinical studies for patient stratification and experimental studies to further elucidate mechanisms involved in the development and progression of CRC.

Acknowledgements

We thank all the clinicians who enrolled patients and participated in the PETACC-3 trial (see Appendix at: http://jco.ascopubs.org/content/27/19/3117.long), in particular the coordinators D Cunningham, R Labianca and E Van Cutsem. We thank F Schütz, T Sengstag and A Wirapati for providing analytical tools for clustering and gene set enrichment analysis. This work was supported in part by the Swiss National Science Foundation (SNF; http://www.snf.ch/E/; Grant No. 320030_135421, to AR and MD); Krebsforschung Schweiz (KFS; http://www.krebsforschung.ch/; Grant No. 02697-08-2010, to AR and MD); Fondation Medic (to EB and MD); and the Swiss National Centre for Competence in Research Molecular Biology (http://www.nccr-oncology.ch/; to VP and MD). ST is a Senior Clinical Investigator of the Fund for Scientific Research Flanders (FWO-Vlaanderen) and has received research grants from the Belgian National Cancer Plan (Nationaal Kankerplan). We thank the TCGA research network for the availability of data.

Author contributions

EB and MD designed the study; YP, FTB, ST, JGH and SW conceived and carried out microarray experiments; YP and FTB performed histopathological experiments and β-catenin scoring; EB, VP, GD, NL and AFN analysed the data, EB, VP, ST, FTB, KOS, NL, JGH, SW, MD and AR performed data interpretation; EB, NL and KOS performed the literature search; and EB generated figures and conceived the first manuscript draft. All authors were involved in writing the paper and had final approval of the submitted and published versions.

References

1. Bosman FT, World Health Organization, International Agency for Research on Cancer. *WHO Classification of Tumours of the Digestive System*, 4th edn. International Agency for Research on Cancer (IARC): Lyons, 2010.
2. Tejpar S, Saridaki Z, Delorenzi M, *et al*. Microsatellite instability, prognosis and drug sensitivity of stage II and III colorectal cancer: more complexity to the puzzle. *J Natl Cancer Inst* 2011; **103**: 841–844.
3. Sinicrope FA, Sargent DJ. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin Cancer Res* 2012; **18**: 1506–1512.
4. Vecchione L, Jacobs B, Normanno N, *et al*. EGFR-targeted therapy. *Exp Cell Res* 2011; **317**: 2765–2771.
5. Martini M, Vecchione L, Siena S, *et al*. Targeted therapies: how personal should we go? *Nat Rev Clin Oncol* 2011; **9**: 87–97.
6. Popovici V, Budinska E, Tejpar S, *et al*. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J Clin Oncol* 2012; **30**: 1288–1295.
7. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* 2007; **50**: 113–130.
8. Shen L, Toyota M, Kondo Y, *et al*. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 2007; **104**: 18654–18659.
9. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* 2008; **10**: 13–27.
10. Furlan D, Carnevali IW, Bernasconi B, *et al*. Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Mod Pathol* 2011; **24**: 126–137.
11. Hinoue T, Weisenberger DJ, Lange CP, *et al*. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012; **22**: 271–282.
12. Loboda A, Nebozhyn MV, Watters JW, *et al*. EMT is the dominant program in human colon cancer. *BMC Med Genom* 2011; **4**: 9.
13. TCGA CGAN. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330–337.
14. Salazar R, Roepman P, Capella G, *et al*. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**: 17–24.
15. Perez Villamil B, Romera Lopez A, Hernandez Prieto S, *et al*. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* 2012; **12**: 260.
16. Wirapati P, Sotiriou C, Kunkel S, *et al*. Meta-analysis of gene expression profiles in breast cancer: toward a unified

understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008; **10**: R65.

17. Farmer P, Bonnefoi H, Becette V, *et al*. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005; **24**: 4660–4671.
18. Shedden K, Taylor JM, Enkemann SA, *et al*. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; **14**: 822–827.
19. Xie T, G DA, Lamb JR, *et al*. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS One* 2012; **7**: e42001.
20. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008; **24**: 719–720.
21. Monti S, Tamayo P, Mesirov J, *et al*. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003; **52**: 91–118.
22. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer: New York, 2009.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; **67**: 301–320.
24. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**: Article 3.
25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
26. Arnholt AT. BSDA: Basic statistics and data analysis. R package v 1.01, 2012; http://cran.r-project.org/web/packages/BSDA/index.html
27. Tejpar S, Bertagnolli M, Bosman F, *et al*. Prognostic and predictive biomarkers in resected colon cancer: current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist* 2010; **15**: 390–404.
28. Mokry M, Hatzis P, de Bruijn E, *et al*. Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One* 2010; **5**: e15092.
29. Hatzis P, van der Flier LG, van Driel MA, *et al*. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* 2008; **28**: 2732–2744.
30. Van der Flier LG, Sabates-Bellver J, Oving I, *et al*. The intestinal Wnt/TCF signature. *Gastroenterology* 2007; **132**: 628–632.
31. Sansom OJ, Reed KR, Hayes AJ, *et al*. Loss of APC *in vivo* immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev* 2004; **18**: 1385–1390.
32. Fevr T, Robine S, Louvard D, *et al*. Wnt/β-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol Cell Biol* 2007; **27**: 7551–7559.
33. de Sousa EMF, Colak S, Buikhuisen J, *et al*. Methylation of cancer stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell* 2011; **9**: 476–485.
34. Merlos-Suarez A, Barriga FM, Jung P, *et al*. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 2011; **8**: 511–524.
35. van der Flier LG, van Gijn ME, Hatzis P, *et al*. Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* 2009; **136**: 903–912.
36. Kosinski C, Stange DE, Xu C, *et al*. Indian hedgehog regulates intestinal stem cell fate through epithelial–mesenchymal interactions during development. *Gastroenterology* 2010; **139**: 893–903.
37. Faris JE, Ryan DP. Trees, forests, and other implications of a *BRAF* mutant gene signature in patients with *BRAF* wild-type disease. *J Clin Oncol* 2012; **30**: 1255–1257.
38. Singh A, Sweeney MF, Yu M, *et al*. TAK1 inhibition promotes apoptosis in *KRAS*-dependent colon cancers. *Cell* 2012; **148**: 639–650.
39. Tanaka H, Deng G, Matsuzaki K, *et al*. *BRAF* mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. *Int J Cancer* 2006; **118**: 2765–2771.
40. Hawkins N, Norrie M, Cheong K, *et al*. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology* 2002; **122**: 1376–1387.
41. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.
42. Tian S, Roepman P, Popovici V, *et al*. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol* 2012; **228**: 586–595.
43. Dahlin AM, Palmqvist R, Henriksson ML, *et al*. The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. *Clin Cancer Res* 2010; **16**: 1845–1855.
44. Barker AD, Sigman CC, Kelloff GJ, *et al*. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther* 2009; **86**: 97–100.
45. Moskaluk CA, Zhang H, Powell SM, *et al*. Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Mod Pathol* 2003; **16**: 913–919.
46. van den Brink GR, Bleuming SA, Hardwick JC, *et al*. Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation. *Nat Genet* 2004; **36**: 277–282.
47. Liu JY, Seno H, Miletic AV, *et al*. Vav proteins are necessary for correct differentiation of mouse cecal and colonic enterocytes. *J Cell Sci* 2009; **122**: 324–334.
48. Zheng H, Ying H, Wiedemeyer R, *et al*. PLAGL2 regulates Wnt signaling to impede differentiation in neural stem cells and gliomas. *Cancer Cell* 2010; **17**: 497–509.
49. Hirose H, Ishii H, Mimori K, *et al*. The significance of PITX2 overexpression in human colorectal cancer. *Ann Surg Oncol* 2011; **18**: 3005–3012.
50. Kontos CK, Papadopoulos IN, Fragoulis EG, *et al*. Quantitative expression analysis and prognostic significance of L-DOPA decarboxylase in colorectal adenocarcinoma. *Br J Cancer* 2010; **102**: 1384–1390.
51. Bhatavdekar J, Patel D, Ghosh N, *et al*. Interrelationship of prolactin and its receptor in carcinoma of colon and rectum: a preliminary report. *J Surg Oncol* 1994; **55**: 246–249.
52. Gaber A, Johansson M, Stenman UH, *et al*. High expression of tumour-associated trypsin inhibitor correlates with liver metastasis and poor prognosis in colorectal cancer. *Br J Cancer* 2009; **100**: 1540–1548.
53. Kawakami K, Ooyama A, Ruszkiewicz A, *et al*. Low expression of gamma-glutamyl hydrolase mRNA in primary colorectal cancer with the CpG island methylator phenotype. *Br J Cancer* 2008; **98**: 1555–1561.
54. Chen Y, Tang Y, Guo C, *et al*. Nuclear receptors in the multidrug resistance through the regulation of drug-metabolizing enzymes and drug transporters. *Biochem Pharmacol* 2012; **83**: 1112–1126.
55. Park SW, Zhen G, Verhaeghe C, *et al*. The protein disulfide isomerase AGR2 is essential for production of intestinal mucus. *Proc Natl Acad Sci USA* 2009; **106**: 6950–6955.
56. Noah TK, Kazanjian A, Whitsett J, *et al*. SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells. *Exp Cell Res* **316**: 452–465.

57. Steppan CM, Brown EJ, Wright CM, *et al*. A family of tissue-specific resistin-like molecules. *Proc Natl Acad Sci USA* 2001; **98**: 502–506.
58. Heiskala K, Giles-Komar J, Heiskala M, *et al*. High expression of RELP (Reg IV) in neoplastic goblet cells of appendiceal mucinous cystadenoma and pseudomyxoma peritonei. *Virchows Arch* 2006; **448**: 295–300.
59. Dalerba P, Kalisky T, Sahoo D, *et al*. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; **29**: 1120–1127.
60. *R Development Core Team. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012.
61. *Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**: R80.
62. *Therenau T. *A Package for Survival Analysis in S*. R package version 2.36–14, 2012.
63. *Bolstad BM, Collin F, Simpson KM, *et al*. Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol* 2004; **60**: 25–58.
64. *Venables WNR, Ripley BD. Modern Applied Statistics with S, 4th edn. Springer: New York, 2002.
65. *Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Statist* 2006; **15**: 651–674.
66. *Van Cutsem E, Labianca R, Bodoky G, *et al*. Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J Clin Oncol* 2009; **27**: 3117–3125.
67. *Jorissen RN, Gibbs P, Christie M, *et al*. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin Cancer Res* 2009; **15**: 7642–7651.
68. *IGC. Expression Project for Oncology, 2008 [cited; available from: http://www.intgen.org/expo/]
69. *Smith JJ, Deane NG, Wu F, *et al*. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010; **138**: 958–968.
70. *Skrzypczak M, Goryca K, Rubel T, *et al*. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One* 2010; **5**: e13091.
71. *Hong Y, Ho KS, Eu KW, *et al*. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007; **13**: 1107–1114.
72. *Gyorffy B, Molnar B, Lage H, *et al*. Evaluation of microarray preprocessing algorithms based on concordance with RT–PCR in clinical samples. *PLoS One* 2009; **4**: e5645.
73. *Galamb O, Sipos F, Solymosi N, *et al*. Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 2835–2845.
74. *Galamb O, Spisak S, Sipos F, *et al*. Reversal of gene expression changes in the colorectal normal–adenoma pathway by NS398 selective COX2 inhibitor. *Br J Cancer* 2010; **102**: 765–773.
75. *Koinuma K, Yamashita Y, Liu W, *et al*. Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene* 2006; **25**: 139–146.
76. *Jorissen RN, Lipton L, Gibbs P, *et al*. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* 2008; **14**: 8061–8069.
77. *Grone J, Lenze D, Jurinovic V, *et al*. Molecular profiles and clinical outcome of stage UICC II colon cancer patients. *Int J Colorectal Dis* 2011; **26**: 847–858.
78. *Birnbaum DJ, Laibe S, Ferrari A, *et al*. Expression profiles in stage II colon cancer according to APC gene status. *Transl Oncol* 2012; **5**: 72–76.
79. *Giancarlo R, Scaturro D, Utro F. Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, *Gap Statistics and Model Explorer. BMC Bioinformat* 2008; **9**: 462.
80. *Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: x1–13.

*Cited only in the Supplementary material.

SUPPLEMENTARY MATERIAL ON THE INTERNET

The following supplementary material may be found in the online version of this article:

Supplementary methods and results (contains a further table and two further figures)

Figure S1. (A) Consensus clustering and similarity dendrogram of samples. (B) Subtype projection in the four-dimensional space of LDA axes. (C) Heat map matrix of pairwise meta-gene Fisher Z-transformed Pearson pairwise correlations. (D) Box plots of intra gene module pairwise gene–gene Pearson correlations in normal samples in both discovery and validation sets

Figure S2. Validation of meta-gene expression pattern of subtypes represented by heat maps

Figure S3. (A) Heat map representing validation of gene expression patterns of subtypes. (B) Pairwise Fisher Z-transformed correlations of meta-genes in validation set. (C) Box plots representing medians of pairwise gene–gene Pearson correlations in the validation datasets

Figure S4. Expression of top five down- and top five u*p* regulated genes from all pairwise comparisons between subtypes

Figure S5. (A) Other clinical and mutational markers tested and found non-significant between subtypes. (B) Clinical variables tested in the clusters of the validation test. (C) Distribution of significant clinical and mutational markers across subtypes. (D) Classification tree trained on clinical variables

Figure S6. Graphs of joined distribution of dominant vsersus secondary patterns in each of the subtypes

Figure S7. Heat map of CNV profiles of 154 samples from the discovery set, randomly ordered inside each of the subtypes

Figure S8. Result of hypothesis testing of median log-scale copy number estimates of chromosome 20 of subtype B versus all other subtypes

Figure S9. Distribution of β-catenin immunoreactivity of the invasion front counts between subtypes

Table S1. Detailed description of gene module members and detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

Table S2. Multiclass linear discriminant (LDA) subtype assignment of samples from validation set

Table S3. Correlations of subtype-specific gene expression profiles (1 versus all moderated *t* test statistics) when accounting for subtype F in the training set

Table S4. Detailed results of meta-gene expression tests pairwise between subtypes and of subtypes to meta-gene medians

Table S5. Detailed results of pairwise comparisons of differentially expressed gene between subtypes

Table S6. Detailed results of Cox proportional hazards models for RFS, OS and SAR for subtype, stage, MSI and *BRAF* and for meta-genes

Table S7. Results of GSEA comparison of enrichment tested signatures in individual subtypes and normal samples

100 Years ago in the *Journal of Pathology*…

The technique of cultivating adult animal tissues *in vitro*, and the characteristics of such cultivations

Albert J. Walton

Experiments on hæmolytic icterus

J. W. M'Nee

Congenital aneurysm in a young rabbit

W. Henwood Harvey

To view these articles, and more, please visit:
www.thejournalofpathology.com

Click 'ALL ISSUES (1892 - 2011)', to read articles going right back to Volume 1, Issue 1.

The Journal of Pathology
Understanding Disease

Journal of
The Pathological Society

[***10***] Belmont PJ, **Budinska E**, Jiang P, Sinnamon MJ, Coffee E, Roper J, Xie T, Rejto PA, Derkits S, Sansom OJ, Delorenzi M, Tejpar S, Hung KE, Martin ES. Cross-species analysis of genetically engineered mouse models of MAPK-driven colorectal cancer identifies hallmarks of the human disease. Dis Model Mech. 2014 Jun;7(6):613-23. doi: 10.1242/dmm.013904. Epub 2014 Apr 17. PMID: 24742783; PMCID: PMC4036469.

© 2014. Published by The Company of Biologists Ltd | Disease Models & Mechanisms (2014) 7, 613-623 doi:10.1242/dmm.013904

RESEARCH ARTICLE

Cross-species analysis of genetically engineered mouse models of MAPK-driven colorectal cancer identifies hallmarks of the human disease

Peter J. Belmont[1,*], Eva Budinska[2,3,*], Ping Jiang[1], Mark J. Sinnamon[4], Erin Coffee[4], Jatin Roper[4], Tao Xie[1], Paul A. Rejto[1], Sahra Derkits[5], Owen J. Sansom[5], Mauro Delorenzi[3], Sabine Tejpar[6], Kenneth E. Hung[7] and Eric S. Martin[1,‡]

ABSTRACT

Effective treatment options for advanced colorectal cancer (CRC) are limited, survival rates are poor and this disease continues to be a leading cause of cancer-related deaths worldwide. Despite being a highly heterogeneous disease, a large subset of individuals with sporadic CRC typically harbor relatively few established ‘driver’ lesions. Here, we describe a collection of genetically engineered mouse models (GEMMs) of sporadic CRC that combine lesions frequently altered in human patients, including well-characterized tumor suppressors and activators of MAPK signaling. Primary tumors from these models were profiled, and individual GEMM tumors segregated into groups based on their genotypes. Unique allelic and genotypic expression signatures were generated from these GEMMs and applied to clinically annotated human CRC patient samples. We provide evidence that a *Kras* signature derived from these GEMMs is capable of distinguishing human tumors harboring *KRAS* mutation, and tracks with poor prognosis in two independent human patient cohorts. Furthermore, the analysis of a panel of human CRC cell lines suggests that high expression of the GEMM *Kras* signature correlates with sensitivity to targeted pathway inhibitors. Together, these findings implicate GEMMs as powerful preclinical tools with the capacity to recapitulate relevant human disease biology, and support the use of genetic signatures generated in these models to facilitate future drug discovery and validation efforts.

KEY WORDS: *KRAS*, *BRAF*, MAPK, Colorectal cancer, GEMM, Genomic signatures

INTRODUCTION

Human sporadic colorectal cancer (CRC) is a complex heterogeneous disease, and this contributes to the low success rate of its clinical trials and lack of robust therapeutics (Betensky et al., 2002; de Bono and Ashworth, 2010). Efforts have been made to understand and account for the heterogeneity of several human cancers, including CRC, with a focus on segmenting cancer populations based on core genetic ‘driver’ lesions (Greenman et al., 2007). In addition, several studies have identified genomic signatures within large CRC datasets that predict clinical outcome (Roth et al., 2010; Dry et al., 2010; Popovici et al., 2012; Budinska et al., 2013; De Sousa E Melo et al., 2013; Sadanandam et al., 2013).

To further understand and experimentally interrogate the biology underlying genetically defined disease segments of interest, and to facilitate discovery of relevant treatment paradigms, stochastic preclinical disease models harboring homologous somatic alterations are crucial. To this end, several studies have utilized genetically engineered model organisms, including *Drosophila* (Vidal and Cagan, 2006; Rudrapatna et al., 2012) and mice (Jonkers and Berns, 2002; Tuveson and Jacks, 2002), to recreate hallmark characteristics of human cancers. *Drosophila* cancer models have shed light on numerous biological underpinnings of cancer, including tumor suppressors, invasion and metastasis (Rudrapatna et al., 2012), providing substrate for further validation in mammalian models. Genetically engineered mouse models (GEMMs) have been utilized as the mammalian cancer model system of choice for decades (Tuveson and Hanahan, 2011; Politi and Pao, 2011). Although GEMMs have traditionally incorporated germline alterations in disease-prevalent genes, models using conditionally controlled, somatically acquired alleles allow a more accurate stochastic modeling of the sporadic nature of human tumorigenesis (Heyer et al., 2010). To address this, GEMMs have been further developed to leverage restricted exposure of Cre recombinase to initiate latent alleles exclusively in tissues of interest, closely mimicking the onset of spontaneous lesions in humans (Johnson et al., 2001; Roper and Hung, 2012; DuPage et al., 2009; Frese and Tuveson, 2007).

To provide maximal experimental utility and enable the translation of preclinical mouse modeling experiments into human disease, GEMMs of human CRC must be driven by homologous allelic series, and exhibit similar clinical presentations to the human disease, including disease histopathology and appearance of metastatic lesions (Heyer et al., 2010; Roper and Hung, 2012). Recently, primary tumors from GEMMs of pancreatic, colorectal and non-small-cell lung cancers harboring genetic lesions that are present in human cancers were shown to be histologically and pathologically similar to their respective human counterparts (DuPage et al., 2009; Hung et al., 2010; Martin et al., 2013). In some cases, GEMMs have closely emulated the response seen in humans to both standard of care and targeted therapies (Arnold et al., 2005); furthermore, the mechanisms of acquired resistance to

[1]Oncology Research Unit, Pfizer Global Research and Development, San Diego, CA 92121, USA. [2]Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, 625 00 Brno, Czech Republic. [3]Bioinformatics Core Facility, SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [4]Division of Gastroenterology, Tufts Medical Center, Boston, MA 02111, USA. [5]The Beatson Institute for Cancer Research, Garscube Estate, Glasgow, G61 1BD, UK. [6]University Hospital Gasthuisberg, Katholieke Universiteit Leuven, 3000 Leuven, Belgium. [7]Pfizer Biotherapeutics Clinical Research, Cambridge, 02140 MA, USA.
*These authors contributed equally to this work

‡Author for correspondence (esmartin.phd@gmail.com)

Received 3 September 2013; Accepted 15 April 2014

TRANSLATIONAL IMPACT

Clinical issue

Colorectal cancer (CRC) is the third leading cause of cancer mortality in the United States, and ~80% of all cases are sporadic in nature, involving the acquisition of tumorigenic somatic alterations. Treatment options for CRC are limited, and the survival rates associated with advanced-stage disease are low. The highly heterogeneous nature of this disease is thought to contribute to the lack of success of novel therapeutics in the clinic. Thus, preclinical models that recapitulate the core biology of the human disease are needed for the identification of new therapeutic strategies. Despite the heterogeneity associated with sporadic CRC, the vast majority of cases display alterations in a limited number of tumor suppressors and oncogenes. Here, the authors amassed a unique collection of genetically engineered mouse models (GEMMs) harboring conditional alleles that mimic acquired somatic alterations observed in human sporadic CRC, including loss of the tumor suppressors *APC* and *TP53* and gain of oncogenic *BRAF* and *KRAS*. To gain an understanding of the utility of these models, gene signatures were derived and used to stratify genomically heterogeneous clinically annotated patient samples, as well as human cell lines treated with targeted inhibitors.

Results

Primary tumors were isolated from GEMMs harboring common CRC 'driver' mutations, and these tumors were subjected to gene expression profiling to generate genotype-specific signatures. GEMM-derived signatures were applied to two independent human clinical CRC datasets for which genomic profiling and survival data were available. The GEMM *Kras* signature score was enriched in individuals with a mutation in *KRAS*, and associated with shorter overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR). Interestingly, the signature further segregated the *KRAS* mutant CRC patient population into two clinically distinct groups, consistent with emerging evidence of heterogeneity in this population in both gene expression and survival. Finally, the signature was predictive of response to MEK inhibitors, which are widely used as cancer drugs, in human CRC cell lines.

Implications and future directions

Together, these results demonstrate that gene signatures derived from genetically and contextually relevant GEMMs are capable of further resolving genomically heterogeneous populations of human CRC and identifying patients with characteristics of aggressive disease. The correlation of the GEMM *Kras* signature with response to targeted inhibition of a clinically relevant pathway in a collection of human CRC cell lines highlights its potential utility in predicting therapeutic response. Future studies will focus on the application of this signature to other therapeutic modalities of interest, and on further understanding the contribution of key nodes or targets present within the signature itself. On a wider scale, this study demonstrates the usefulness of GEMMs expressing conditional alleles for exploring genetic heterogeneity in human malignancies.

such agents have often closely resembled those seen in the clinic (Engelman et al., 2008; Jorissen et al., 2009; Van Cutsem et al., 2009; Hegde et al., 2013). Thus, GEMMs are useful preclinical models for modeling human cancer biology and identifying potential therapeutic targets.

To further our understanding of the molecular etiology underlying common genotypic subsets of human CRC, and to assess the extent to which they recapitulate human disease in animal models, we amassed a collection of GEMMs that combine colon-specific mutations, including somatic alterations in *Apc* (*Apc*$^{\mathrm{CKO}}$), *Tp53* (*Tp53*$^{\mathrm{flox/flox}}$), *Kras* (*Kras*$^{\mathrm{LSL\text{-}G12D}}$) and *Braf* (*Braf*$^{\mathrm{V600E}}$), genes that are among the most frequently mutated in human sporadic CRC (Cancer Genome Atlas Network, 2012). Primary tumor material from this collection was subjected to gene expression profiling to assess core similarities and differences among these models, and to generate unique signatures based on genotype. These signatures were then evaluated in human CRC tissue with annotated clinical data to assess the ability of these GEMMs to recapitulate the core transcriptional biology of their human CRC counterparts. Overlapping gene expression modules shared between GEMM and human signatures represent potential points of therapeutic interrogation and provide key substrate for follow-up validation and drug discovery efforts.

RESULTS

Development and profiling of genetically relevant CRC GEMMs

Adult GEMMs harboring combinations of latent, inactive alleles of the four most common somatic lesions observed in human CRC (Cancer Genome Atlas Network, 2012) (*APC*, *TP53*, *KRAS* and *BRAF*) were subjected to surgically restricted delivery of *AdCre* to the distal colon; mice were then followed longitudinally for tumor progression via endoscopy, and tumor material was harvested as previously described (Hung et al., 2010; Martin et al., 2013). The conditional *Apc* and *Tp53* alleles harbor *loxP* sites (floxed), which, upon exposure to *AdCre*, result in excision of critical exons, resulting in loss-of-function proteins, as previously described (Kuraguchi et al., 2006; Kirsch et al., 2007). The conditional *Kras* and *Braf* alleles harbor floxed transcriptional stop elements upstream of mutant forms of exon 1 (*Kras*$^{\mathrm{G12D}}$) (Hung et al., 2010) or exon 15 (*Braf*$^{\mathrm{V600E}}$) (Coffee et al., 2013). A list of primary tumors with allelic combinations is provided (supplementary material Table S1). Tumors and normal colonic tissue from wild-type littermate controls were subjected to whole-genome expression profiling. Subsequently, principal component analysis (PCA) and unsupervised hierarchical clustering on the top 500 most variable genes was performed. Individual CRC GEMMs clustered by genotype, both in the PCA (Fig. 1A, genotype representing the first two principal components) and hierarchical clustering (Fig. 1B). These results demonstrate that the genotypes of these models represent the primary differentiating feature, and suggest that each genotype likely possesses unique underlying biological characteristics.

Allele-specific GEMM signatures

To further assess the underlying differences among our CRC models, we identified gene signatures (lists of differentially expressed genes) characteristic of each mutant allele (*Apc*, *Tp53*, *Kras*, *Braf*) within the GEMM collection using a multivariable analysis (see Materials and Methods). It is important to note that all GEMMs contain *Apc* lesions; therefore, all results for *Braf*, *Kras* and *Tp53* alleles should be interpreted with this regard. A Venn diagram (Fig. 2A) and heatmaps of supervised hierarchical clustering on the signature-specific genes (Fig. 2B-E) demonstrate that these gene lists partially overlap, suggesting common biological characteristics, including redundant signaling and pathway activation. To determine whether the unique or intersecting gene lists associated with each mutant allele displayed enrichment in known biological processes or curated gene signatures, we cross-referenced each to the molecular signatures database [MSigDB (www.broadinstitute.org/gsea/msigdb/)]. Indeed, common gene sets enriched among upregulated *Kras* and *Braf* genes included several annotated MAPK pathway sets, consistent with the established roles of mutant *Kras* and *Braf* in activating this pathway (supplementary material Table S2). Gene sets enriched among shared upregulated *Apc* and *Tp53* genes included several cell cycle gene sets as well as DNA synthesis, replication and repair, consistent with their established roles as tumor suppressors and thus with the deregulation of these functions in our models (supplementary

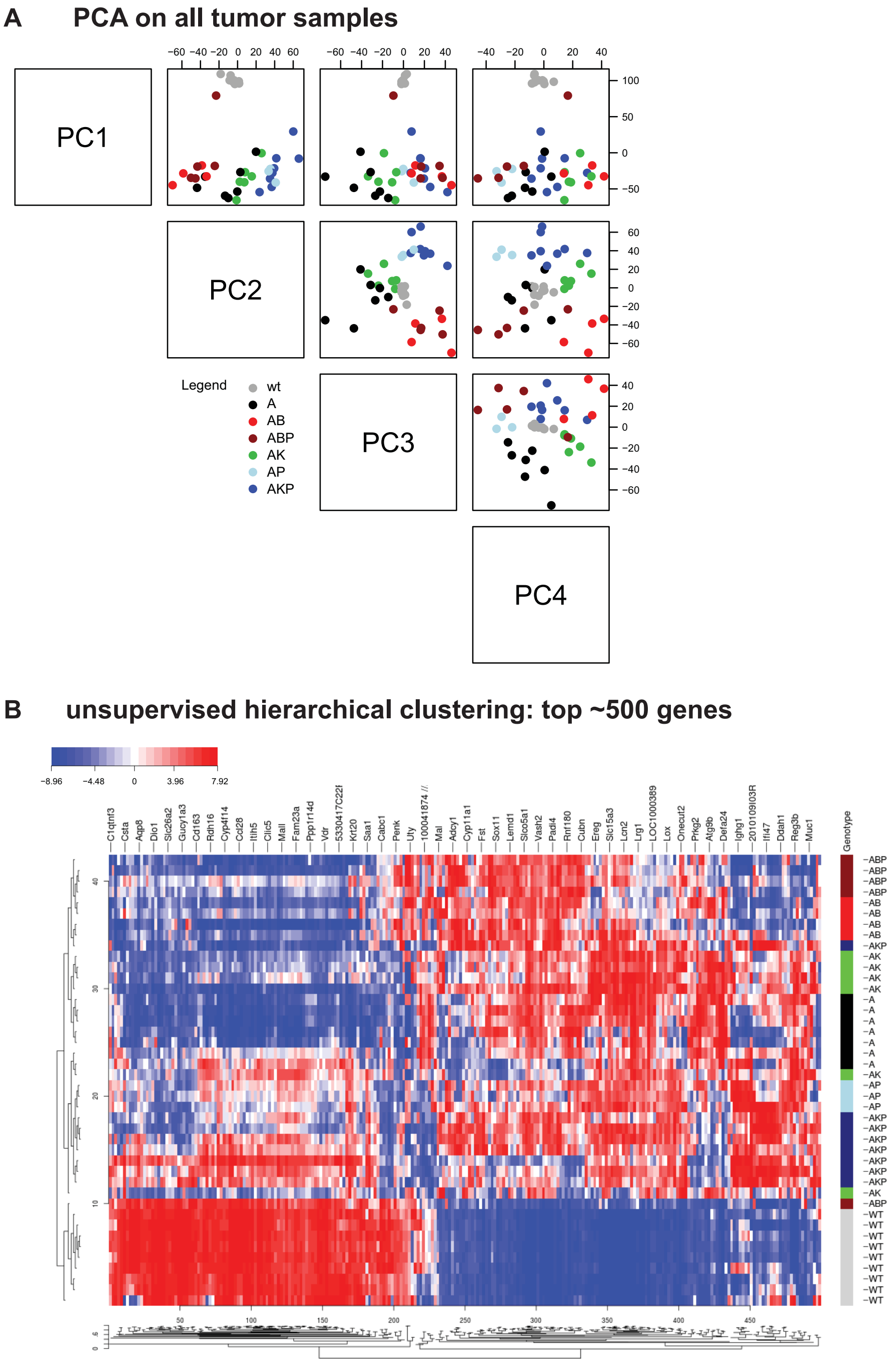


Fig. 1. GEMM primary tumors segregate by genotype. (A) Principal component analysis (PCA) on GEMM primary tumor samples and normal colon tissue. wt, normal colon from wild-type untreated mice. The following designations describe the alleles present in CRC GEMMs: A, *Apc*; AB, *Apc*, *Braf*; ABP, *Apc*, *Braf*, *Tp53*; AK, *Apc*, *Kras*; AKP, *Apc*, *Kras*, *Tp53*; AP, *Apc*, *Tp53*. (B) Unsupervised hierarchical clustering using the top 500 most differentially expressed genes from samples as described in A. WT, wild type.

material Table S3). Gene sets enriched among unique genes for each allele were also assessed. Gene sets found to be enriched in *Kras*-specific genes included metabolism, signaling downstream of receptors, and adhesion (supplementary material Table S4), functions previously ascribed to mutant *KRAS* (Racker et al., 1985; Pollock et al., 2005; Rajalingam et al., 2007; Levine and Puzio-Kuter, 2010). Interestingly, gene sets enriched among unique *Braf* genes also include metabolism, consistent with previously established links between oncogenic *BRAF* and metabolic deregulation (Yun et al., 2009); however, additional gene sets included immune response signaling, consistent with additional roles for oncogenic *BRAF* (Sumimoto et al., 2006) (supplementary material Table S5). Gene sets found to be enriched in *Apc*-specific genes included development (supplementary material Table S6), consistent with the role of aberrant *APC* in WNT–β-catenin signaling and development (Clevers, 2006), as well as several gene sets associated with small-molecule transport, a role to our knowledge not fully characterized for aberrant *APC*. Gene sets enriched in *Tp53*-specific genes included ubiquitylation and proteolysis pathways (supplementary material Table S7), consistent with the central role of these pathways in regulating endogenous *TP53* (Lee and Gu, 2010). Taken together, these findings indicate that lesions in our GEMM alleles of interest result in gene signatures characteristic of known or putative biological roles for each allele.

Generation and validation of GEMM allelic signatures

We defined GEMM allele-specific scores as a difference of average gene expression between the top 100 up- and top 100 downregulated

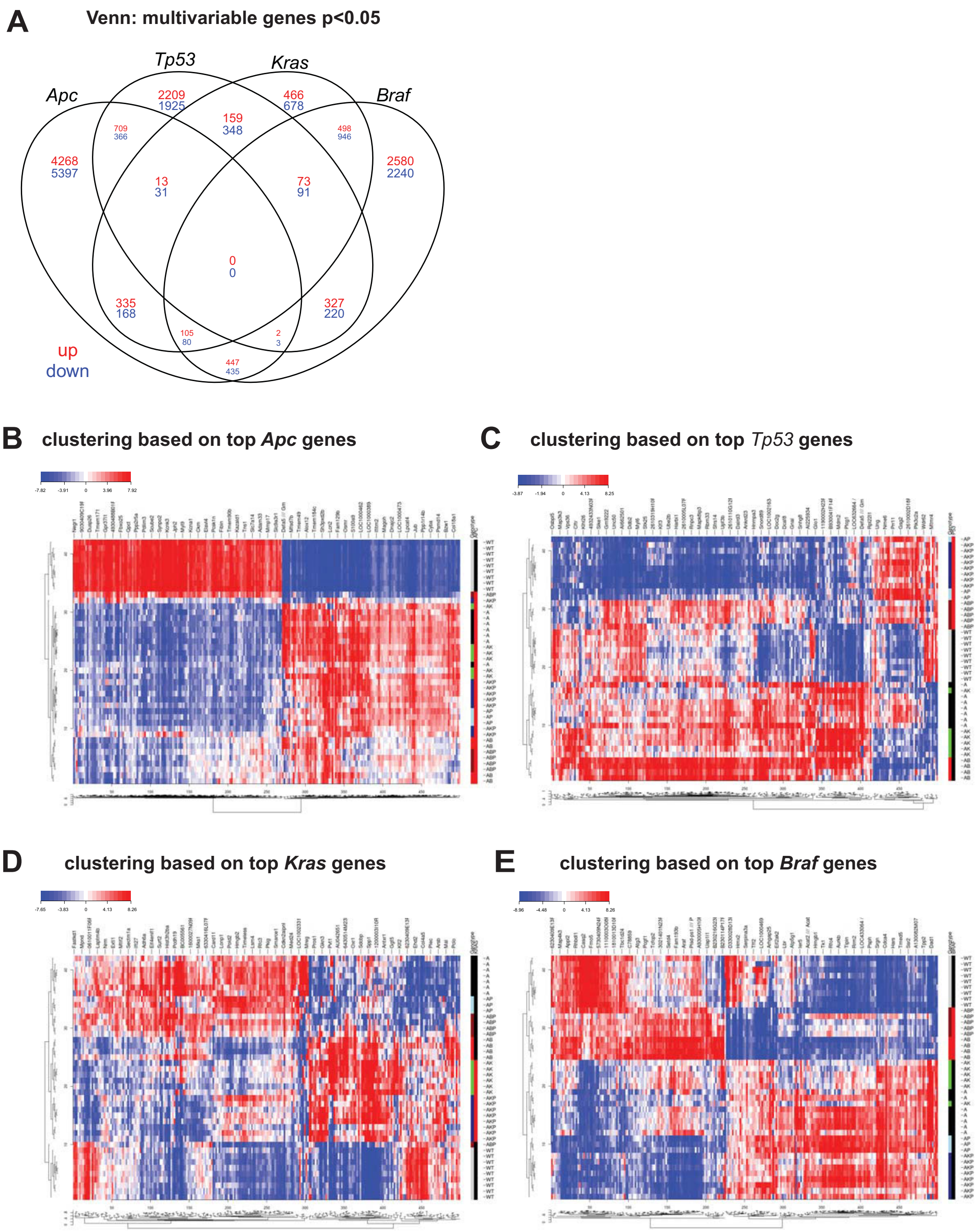


Fig. 2. Multivariable analysis identifies genes associated with each allele from the GEMM cohort. (A) Venn diagram depicting the number of unique or shared genes associated with each GEMM allele. Red, upregulated genes; blue, downregulated genes. (B-E) Clustering of GEMMs based on expression profiles of genes associated with each allele.

genes from the corresponding signature. The score for each individual GEMM allelic signature (*Kras*, *Braf*, *Apc*, *Tp53*; supplementary material Tables S8-S11, respectively) was computed in each of the models (A: *Apc*; AK: *Apc*, *Kras*; AKP: *Apc*, *Kras*, *Tp53*; AB: *Apc*, *Braf*; ABP: *Apc*, *Braf*, *Tp53*; AP: *Apc*, *Tp53*; WT, wild type; supplementary material Table S1). As expected, the models containing a given mutation had the highest score for that allelic signature in the discovery set (Fig. 3A-D). For instance, the GEMM *Apc* signature score was high in all GEMM models, because all models contain this mutation (Fig. 3A), whereas the GEMM *Tp53* signature was high in models containing *Tp53*, including AP, ABP and AKP, but low in A, AB and AK (Fig. 3B). In the case of the GEMM *Kras* signature, the score was high in models containing *Kras*, including AK and AKP (Fig. 3C). The highest *Braf* score was found in models containing *Braf*, including AB and ABP (Fig. 3D). Interestingly, the GEMM *Kras* score was also high in models with *Braf* and *Apc* mutation (AB), but not in those containing *Braf*, *Apc* and *Tp53* mutation (ABP) (Fig. 3C), suggesting that the addition of *Tp53* to the *Apc*, *Braf* mutant background might result in less reliance on MAPK-driven signaling. Similar trends were seen in other genotypes, with *Tp53* mutation leading to a systematically lower signature score compared with their counterparts without the

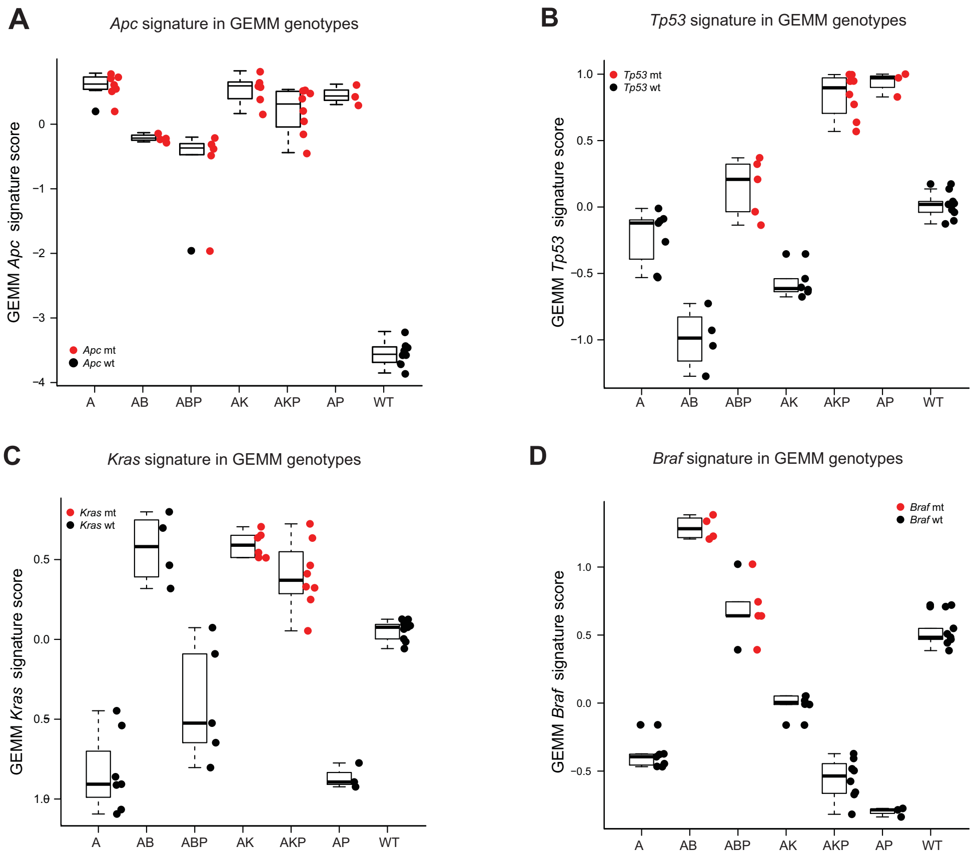


Fig. 3. Internal validation of the GEMM signatures to distinguish models containing mutant alleles. (A-D) GEMM allele-specific signatures were generated (*Apc*, *Tp53*, *Kras*, *Braf*), and a signature score was calculated in each of the six GEMM genotypic models as well as WT normal colon control.

mutation (*Apc* signature in AP versus A, Fig. 3A; *Kras* signature in AKP versus AK, Fig. 3C; ABP versus AB, Fig. 3D). A potential explanation for these observations could include the increased presence of genomic instability, a well-known consequence of aberrant *Tp53*.

We next applied the signature to an independent GEMM CRC sample set consisting of acute activation of shared alleles, including *Apc*, *Tp53* and *Kras*. Consistent with the findings in our discovery cohort, our GEMM allelic signatures scored highest in GEMMs derived from an independent cohort that contained the corresponding mutant allele (supplementary material Fig. S1A-C), further validating their predictive utility.

Overlap of allele-specific GEMM *Kras* and *Braf* signatures with clinically annotated CRC datasets

To assess the extent to which our GEMMs recapitulate the genetic and biological features of human CRC, and to assess the utility of this collection for preclinical studies, we compared their genomic signatures to those of clinically annotated human CRC datasets. To this end, we utilized the Pan-European Trials in Alimentary Tract Colon Cancers (PETACC-3), a large Phase III randomized trial in which 688 patients with stage II or III CRC were characterized by genomic and mutational analysis, including *KRAS* and *BRAF*. Because the mutant *Kras* allele in the GEMM cohort (*Kras*$^{\mathrm{LSL\text{-}G12D}}$) is a gain-of-function mutation, for the purpose of comparison we considered all *KRAS* gain-of-function mutations in the PETACC-3 dataset, with the caveat that different types of *KRAS* mutations potentially have unique biological characteristics (Kirk, 2011). As indicated in Fig. 4A, the average GEMM *Kras* signature score was significantly higher in patients with the *KRAS* mutant than those with wild-type *KRAS*. Given the variability in the GEMM *Kras* signature score among individuals with wild-type *KRAS* and the fact that our *Kras* signature scored high in our *Braf*-containing models, possibly picking up on common MAPK pathway mechanisms, these patients were further annotated based on *BRAF* mutation or similarity to a published *BRAF*-like signature (Popovici et al., 2012). Interestingly, of the *KRAS* wild-type patients, both *BRAF* mutant (Fig. 4A, red circles) as well as those with a high *BRAF*-like signature score (Fig. 4A, green circles) tended to display a higher signature score, supporting our hypothesis that, in addition to distinguishing *KRAS* mutant patients, the GEMM *Kras* signature also captures those with high MAPK pathway activity. Together, these data indicate that the GEMM signature is enriched in patients with *KRAS* mutation, as well as *BRAF* mutation or a high degree of similarity to a *BRAF*-like signature, the latter of which is potentially indicative of a common biology shared among *KRAS* and *BRAF* mutant patients.

To determine whether our GEMM *Kras* signature is representative of human *KRAS* mutant CRC tumors, we compared it to a human *KRAS* signature derived in the multivariable model with *KRAS* and *BRAF* mutation as covariates in PETACC-3 patients. Consistent with the GEMM, the PETACC-3 *KRAS* signature score was higher among *KRAS* mutant patients than *KRAS* wild-type patients, whereas, again, *BRAF* mutant and *BRAF*-like patients tended to score highest among the *KRAS* wild-type population (Fig. 4B). The GEMM and PETACC-3 *KRAS* signature scores showed a high degree of correlation both among GEMMs (Fig. 4C, R^2=0.74) and among patients (Fig. 4D, R^2=0.32). These findings suggest that the *Kras* signature derived from a relatively homogeneous background such as the GEMM might be capable of capturing common and disease-relevant biology present in human *KRAS* patients.

Interestingly, our GEMM *Braf* signature score did not correlate with the human *BRAF* signature score of Popovici et al. (Popovici

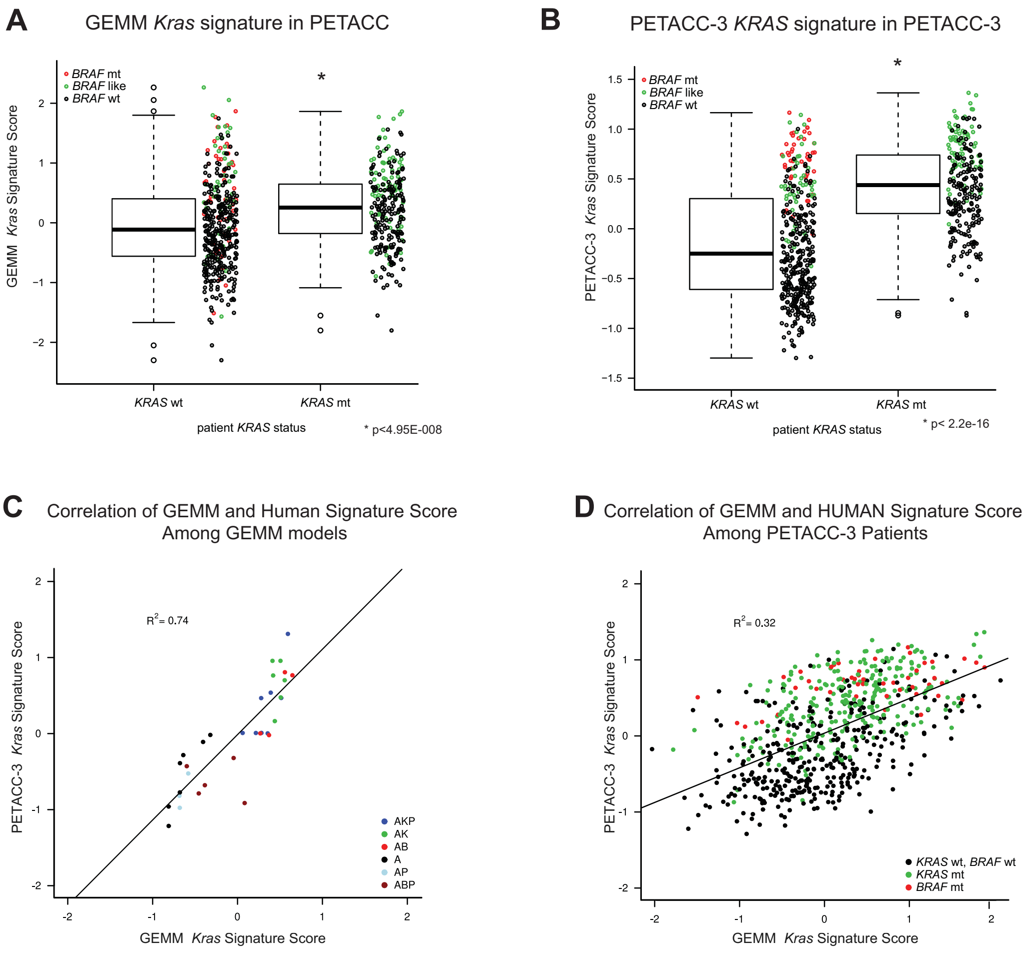


Fig. 4. GEMM *Kras* signature can distinguish *KRAS* mutant patients. (A) Box plot depicting distribution of IQR normalized GEMM *Kras* signature score in PETACC-3 patients that are *KRAS* wild-type (wt) or mutant (mt). *KRAS* wild-type patients were further annotated as *BRAF* mutant (red) or containing significant enrichment of a *BRAF*-like signature (Popovici et al., 2012) (green). (B) Box plot depicting distribution of IQR normalized PETACC-3 *KRAS* signature score in PETACC-3 patients. *KRAS* wild-type patients were further annotated as in A. (C,D) Correlation between IQR normalized *KRAS* scores derived from PETACC-3 (*x*-axis) or GEMM (*y*-axis) signatures among the GEMM collection (C) and PETACC-3 patients (D). Solid line represents linear fit. Colors in C represent each GEMM genotype. Colors in D represent status of *KRAS* and *BRAF* mutation in PETACC-3 patients: red, *BRAF* mutant; green, *KRAS* mutant.

et al., 2012), nor was it able to predict *BRAF* mutant tumors in the PETACC-3 data. Also, the recent *BRAF* signature derived from human samples did not predict correctly *Braf* mutant status in our GEMMs (data not shown). This, together with the results of the *Braf* signature pathway analysis pointing to proliferation, shows that our *Apc*-based *Braf* models are potentially less representative of the human *BRAF* mutant population. This is consistent with the low frequency of concomitant *BRAF* and *APC* lesions observed in human cases (Cancer Genome Atlas Network, 2012).

Clinical characteristics of patient samples based on GEMM *Kras* signature score

We assessed differences in available clinical variables among all individuals in the PETACC-3 cohort. Patient populations were defined based on each GEMM signature score into allele-like and non-allele-like groups (threshold 0 on inter-quartile range normalized scores). GEMM *Kras*-like tumors exhibited a statistically significant enrichment for various characteristics, including mucinous histology, *KRAS* mutant, *BRAF* mutant, right-side, stage 3, and similarity to a *BRAF*-like population shown previously to be associated with poor prognosis (Popovici et al., 2012) (supplementary material Table S12), implicating the ability of the GEMM *Kras* signature at distinguishing aspects of advanced disease.

GEMM *Kras* signature is associated with poor outcome

To determine whether the GEMMs are representative of advanced disease, we examined survival differences among annotated patients in PETACC-3. Differences in overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR) were compared. To validate our findings, we performed a similar assessment on an independent publicly available sample cohort (GEO GSE14333) (Jorissen et al., 2009), consisting of 115 stage II/III human CRC samples with gene expression profiling and survival data. Of the four core GEMM signatures generated (*Apc*, *Tp53*, *Braf*, *Kras*), the *Kras* signature score produced the highest hazard ratios for OS and SAR in the PETACC-3 dataset, and among the highest hazard ratios for OS, RFS and SAR in the GSE14333 dataset (Table 1), suggesting that it is most indicative of advanced disease. OS, RFS and SAR based on GEMM *Kras* signature was plotted for the PETACC-3 dataset (Fig. 5A-C) and for the GSE1433 dataset (Fig. 5D-F). Additional Kaplan-Meier plots for GEMM *Braf*, *Apc* and *Tp53* signatures in PETACC-3 as well as GSE144333 can be found in supplementary material Figs S2 and S3, respectively. Because the GEMM *Kras* signature was associated with some prognostic clinical variables (e.g. stage), we also fitted a multivariable survival model with GEMM *Kras*-like signature, *BRAF* mutant, *KRAS* mutant, mucinous status, grade and MSI, within stage-3 patients of the PETACC-3 dataset (stage 2 patients were enriched for relapsed patients, so were not representative of the population). The GEMM *Kras* signature remained significant for both OS and RFS (supplementary material Table S13). Together, these findings suggest that our GEMM *Kras* signature could offer insight into survival characteristics in two independent large human CRC patient cohorts.

Given that *KRAS* mutant CRC patients have been shown to be heterogeneous (Budinska et al., 2013; Sadanandam et al., 2013) and

Table 1. Survival characteristics associated with each GEMM signature

	PETACC-3		GSE14333	
Parameter	HR	*P*-value	HR	*P*-value
Kras-like vs non *Kras*-like				
OS	1.64	0.00077	2.72	0.00656
RFS	1.46	0.00251	3.25	0.00132
SAR	1.49	0.01204	4.28	0.01616
Braf-like vs non *Braf*-like				
OS	1.58	0.00142	0.88	0.71205
RFS	1.72	0.00001	1.54	0.22355
SAR	0.9	0.48413	0.94	0.89929
Tp53-like vs non *Tp53*-like				
OS	0.64	0.00144	0.93	0.84505
RFS	0.59	0.00001	0.31	0.00128
SAR	1.1	0.55328	1.08	0.88514
Apc-like vs non *Apc*-like				
OS	0.73	0.02836	2.72	0.01102
RFS	0.75	0.01871	1.45	0.28122
SAR	0.94	0.68965	1.71	0.30573

GEMM *Apc*, *Tp53*, *Kras* and *Braf* signatures were applied to the PETACC-3 and GSE14333 datasets as described in Fig. 5, and OS, RFS and SAR were compared for each respective signature. Shown are *P*-values and hazard ratios (HR) for each parameter.

given the ability of the GEMM *Kras* signature to distinguish patients with poor prognosis, we sought to determine whether this signature could further delineate clinical features, specifically in a *KRAS* mutant patient population. Although not statistically significant, a trend toward worse prognosis was observed for *KRAS* mutant patients with high GEMM *Kras* signature score for OS, RFS and SAR (Fig. 6A-C, P=0.480, P=0.398 and P=0.341, respectively).

Together, these data indicate that the GEMM *Kras* signature can distinguish a subpopulation of patients with poor prognosis, perhaps owing to its ability to further distill a heterogeneous patient population to the core underlying biology beyond simply the status of a given driver lesion, much like the recent *BRAF* signature (Popovici et al., 2012) with which it is correlated.

GEMM *Kras* signature is predictive of sensitivity to targeted inhibitors

To determine the utility of the GEMM *Kras* signature as a preclinical model selection tool, we assessed its ability to predict response to targeted inhibitors in a panel of cell lines. Given the clinical potential in applying MEK inhibitors to treat various tumor types, including CRC, we sought to determine whether the GEMM signature was predictive of response to these inhibitors as determined by a publicly available study of drug sensitivity across a comprehensive collection of cancer cell lines (http://www.cancerrxgene.org), with a focus on CRC. A high GEMM *Kras* signature score was associated with increased sensitivity of CRC cell lines to two independent MEK inhibitors used in the study, PD-0325901 and AZD6244 (Fig. 7A,B, respectively). To independently validate these findings, we selected representative cell lines with relatively high and low GEMM *Kras* signature scores (high: LS-1034, LS-513; low: Colo-320, SW948), and assessed cell viability following a full-dose response of these MEK inhibitors. The cell lines with higher GEMM *Kras* signatures displayed relatively greater sensitivity than those lines with lower GEMM *Kras* signatures to the MEK inhibitors PD-0325901 and AZD6244 (Fig. 7C,D, respectively). This supports our hypothesis that the GEMM *Kras* signature is associated with an increased dependency on MAPK signaling, and therefore an enhanced sensitivity to pathway inhibition via selective targeting of MEK. This is consistent with the known 'driver' phenotype of mutant *KRAS* and the increased dependency on the MAPK pathway observed in several *KRAS* mutant cell lines. Interestingly, the GEMM *Kras* signature score added predictive utility beyond simply *KRAS* mutation status of the cell lines: a signature score positively correlated with sensitivity to MEK inhibition, even within a set of *KRAS* mutant cell lines. Taken together, these findings provide

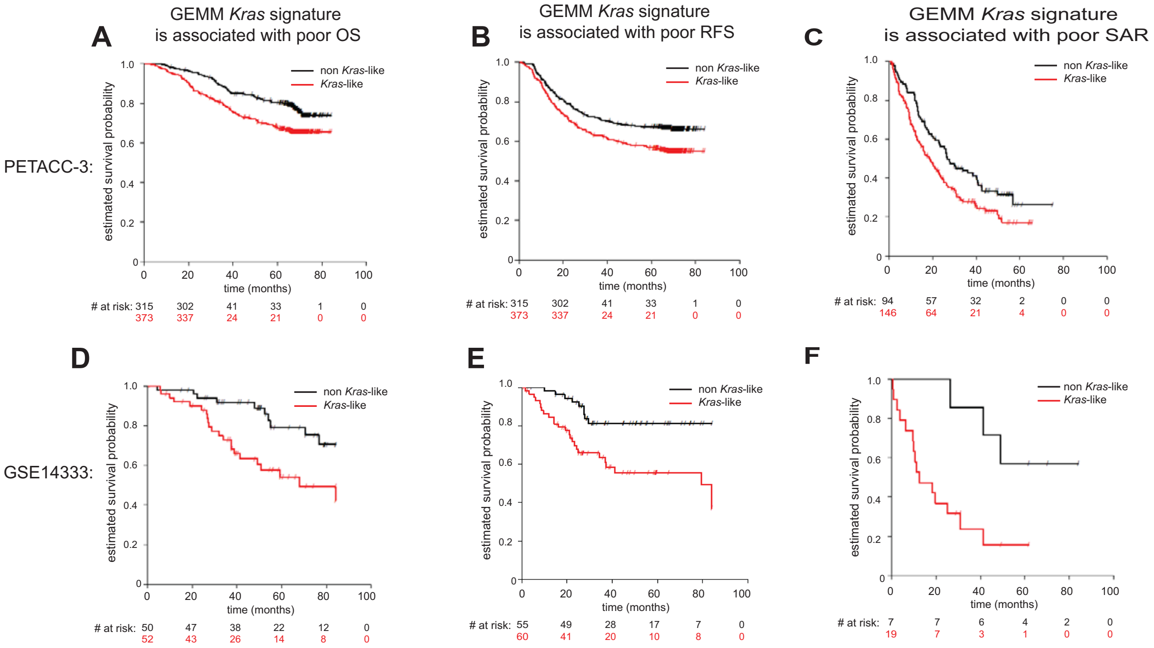


Fig. 5. GEMM *Kras* signature is predictive of poor prognosis in two independent clinically annotated CRC patient datasets. (A-C) Kaplan-Meier estimates of three survival end points among PETACC-3 patients (OS, RFS and SAR) between *Kras*-like (red) and non *Kras*-like (black) groups, as defined according to the *Kras* gene expression signature derived from GEMMs. (D-F) Same as A-C but for patients within the GSE14333 dataset. Survival times were cut at 84 months.

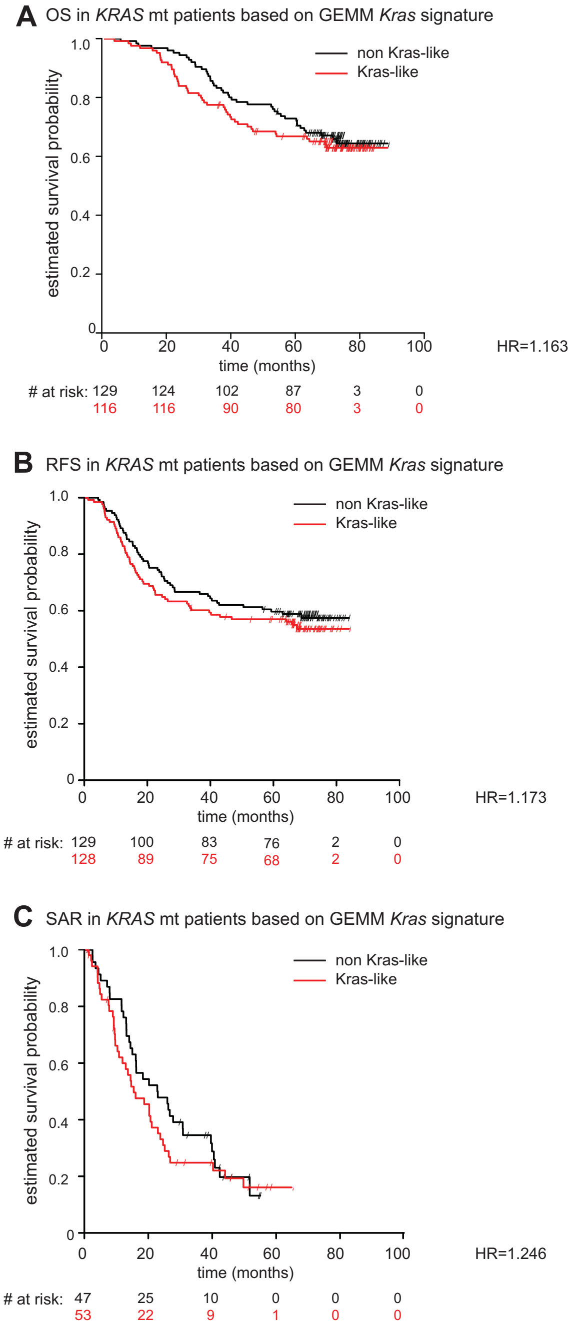


Fig. 6. GEMM *Kras* similarity score assigned within the PETACC-3 *KRAS* mutant population is associated with poor prognosis. (A-C) The PETACC-3 *KRAS* mutant population was treated separately and a similarity score was assigned to each *KRAS* mutant patient, based on similarity to the GEMM *Kras* signature. Kaplan-Meier curves demonstrating that the GEMM *Kras* signature is predictive of poor overall survival (OS, A), relapse-free survival (RFS, B), and survival after relapse (SAR, C) within the PETACC-3 *KRAS* mutant population. Survival times were cut at 84 months.

motivation for using the GEMM *Kras* signature for predicting response to targeted inhibitors of the MAPK pathway, including those targeting MEK.

DISCUSSION

The identification of core ‘driver’ lesions among tumor indications provides a means for segmenting patients and, in some cases, selecting treatment regimens. Despite advances in patient stratification and treatment selection, there are still sizeable segments of human disease with limited effective treatment options. One such segment is defined by the presence of *KRAS* mutations, constituting roughly 30-40% of sporadic CRC (Jorissen et al., 2009; Cancer Genome Atlas Network, 2012). Further compounding this problem is the lack of informative preclinical models in which to conduct rapid drug discovery efforts.

Next-generation GEMMs have gained prominence as preclinical cancer models (DuPage et al., 2009; Heyer et al., 2010; Politi and Pao, 2011). Specific advantages of these models include the ability to selectively activate latent alleles of interest, effectively modeling the stochastic gain of activating mutations and/or loss of tumor suppressors commonly observed in sporadic human cancers. Our GEMM collection contains combinations of genes frequently mutated or lost in human CRC, including *Apc*, *Tp53*, *Braf* and *Kras*, thereby allowing us to model a broad spectrum of human disease. Adding to the utility of these models, primary tumors are used as substrate to generate tumor-derived cell lines that maintain much of the biology of the original tumors, and retain key alleles of interest (Martin et al., 2013). Further, these cell lines serve as a platform for *in vitro* and *in vivo* interrogation because they are amenable to growth in subcutaneous space, in sites common for metastasis such as the liver, and in the native colonic environment of syngeneic, immunocompetent recipients (Martin et al., 2013). As in any GEMM, there are also clear drawbacks to these models, such as the limited number of defined genetic lesions and tumor heterogeneity relative to their human counterparts, in large part due to the inherent nature of an inbred model. In addition, owing to their historically short lifetime as preclinical models, their translational value of has yet to be fully realized. Thus, it is important to understand the role of these models as a complementary tool in a larger comprehensive preclinical drug discovery program.

In the current study, we investigated the genomic characteristics of primary tumors from our collection of CRC GEMMs containing genetic lesions that are present in a large portion of human disease cases. The genomic profiles of these tumors properly segregated based on their core genotypes, with each genotype containing unique distinguishing signatures. Our *Braf* models were exclusively generated along with loss of *Apc*, a condition likely not indicative of human CRC progression as indicated by a recent assessment of human CRC mutational data (Cancer Genome Atlas Network, 2012) and also reflected in our GEMM *Braf* signature failing to classify *BRAF* mutant clinical samples.

The GEMM *Kras* signature was effectively validated within an independent collection of GEMMs, as it properly distinguished *Kras* mutant models from non-mutant. A more detailed analysis of the GEMM *Kras* signature revealed that it was enriched in human CRC patients with advanced disease and poor prognosis. The signature was also able to further stratify the *KRAS* mutant segment of a large clinical cohort, suggesting that a comprehensive signature can provide additional power in further segregating a patient population of interest, beyond simply the status of a given driver lesion, and indicating that there are likely additional underlying characteristics that account for severity of disease beyond the mutation status of *KRAS*. Finally, the signature provided additional utility in predicting sensitivity to targeted MEK inhibition across a panel of CRC cell lines, because those lines with a high signature score tended to display increased sensitivity to two independent MEK inhibitors, suggesting a utility in predicting pathway dependence. The correlation was maintained even within a set of cell lines that harbor *KRAS* mutation: *KRAS* mutant cell lines with relatively higher signature scores displayed increased sensitivity compared with mutant lines with lower signature scores.

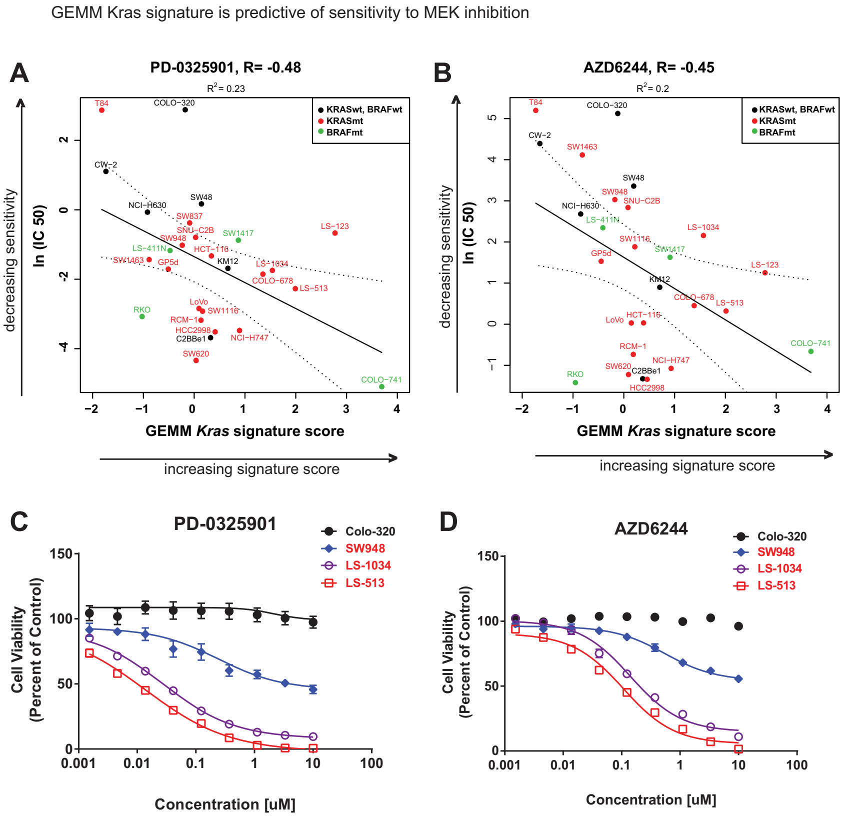


Fig. 7. GEMM *Kras* signature predicts sensitivity to targeted MEK inhibition. The top 100 most significant GEMM *Kras* signature genes were used to segregate cell lines based on similarity to these genes (*x*-axis, GEMM *Kras* signature score; increasing value indicates increasing signature score), and compare this to the relative sensitivity to targeted MEK inhibitors reported in the Sanger dataset (www.cancerrxgene.org) [*y*-axis, ln (IC_{50}); increasing value indicates decreasing sensitivity to the inhibitor], including PD-0325901 (A) and AZD6244 (B). *KRAS* mutant cell lines are in red, *BRAF* mutant cell lines are in green, *KRAS*/*BRAF* wild-type cell lines are in black. The associated Pearson correlations and R^2 values relating *Kras* signature score to inhibitor sensitivity are shown above each graph. (C,D) Independent confirmation of sensitivity to MEK inhibitors. Representative cell lines with relatively high (LS-1034, LS-513) and low (Colo-320, SW948) GEMM *Kras* signature scores were experimentally tested as an independent assessment of sensitivity to MEK inhibitors PD-0325901 (C) and AZD6244 (D). *KRAS* mutant cell line names are in red, *KRAS*/*BRAF* wild-type cell line names are in black.

This approach could potentially be used to identify additional pathway dependencies and corresponding therapeutic sensitivities. Taken together, this study highlights instances in which signatures generated from the GEMMs are applicable to recapitulating biological characteristics of human disease, including prognosis and response to targeted therapeutics. Although several limitations preclude the use of GEMMs as a stand-alone discovery model, the features described herein provide further insight into the power of these GEMMs of sporadic CRC as a companion preclinical discovery model in a comprehensive drug discovery effort.

MATERIALS AND METHODS

This research protocol was approved by our attending veterinarian, and by the Pfizer Institutional Animal Care and Use Committee (IACUC).

CRC GEMMs

The generation and genotyping of *Apc* (*Apc*^{CKO}), *Tp53* (*Tp53*$^{flox/flox}$), *Kras* (*Kras*$^{LSL-G12D}$) and *Braf* (*Braf*^{V600E}) genetically engineered mice has been previously described (Hung et al., 2010).

CRC GEMM tumor samples and gene expression data

Murine primary tumor samples from GEMMs treated with AdCre, and normal colon tissue from untreated wild-type mice were collected. Wild-type mouse colon tissue used for RNA extraction and microarray analysis was enriched for epithelial cells. Briefly, colons were opened lengthwise, cut into 3-5 mm fragments, and washed in HBSS-glucose. Fragments were then resuspended in 20 ml HBSS-glucose-dispase-collagenase solution, placed into a conical tube and agitated on a shaking platform for 25 minutes at 25°C. The digested tissue was further disaggregated by hand pipetting and vigorous shaking for 3 minutes and inspected under an inverted microscope. Subsequently, enzymes were neutralized with 50 ml DMEM-sorbitol and crypt cell suspensions were separated from intestinal fragments and passed through a 70-µm cell strainer. The epithelial-enriched fraction was briefly centrifuged and used for RNA extraction and microarray analysis. RNA was isolated and processed for hybridization on Mouse Affymetrix GeneChip 430 2.0 arrays (Affymetrix, Santa Clara, CA) as previously described (Martin et al., 2013). All gene expression data can be found at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE50794. Our training set consisted of Affymetrix Mouse 430 2.0 gene expression profiles of 33 primary tumors representing the following genotypes: *Apc* (7), *Apc*/*Kras* (6), *Apc*/*Kras*/*Tp53* (8), *Apc*/*Tp53* (3), *Apc*/*Braf* (4), *Apc*/*Braf*/*Tp53* (5) and nine normal colon tissue samples.

The validation set consisted of Affymetrix Mouse 430 2.0 gene expression profiles of 15 primary tumors of genotypes: *Apc* (3), *Apc/Kras* (6), *Apc/Tp53*(6) and three normal colon tissues.

Clinical and cell line data

803 stage II or III human CRC gene expression profiles from both the PETACC-3 trial [688 formalin-fixed paraffin-embedded samples profiled on ALMAC CRC DSA platform (Almac, Craigavon, UK) (Budinska et al., 2013)] and Moffit samples [115 fresh frozen samples profiled on Affymetrix HG U133+ 2.0 platform (Jorissen et al., 2009)] with available clinical and survival data were used to test whether our GEMM models are representative of human disease. The PETACC-3 data are available from the Array Express database under the accession number E-MTAB-990; the Moffit data are available from the GEO database under accession number GSE14333. Cell line gene expression profiles with drug sensitivity (http://www.cancerrxgene.org) (Garnett et al., 2012) profiled on Affymetrix HG U133A platform (Affymetrix, Santa Clara, CA) were downloaded from the Array Express database under the accession number E-MTAB-783.

Microarray data normalization and data filtering

All Affymetrix gene expression data were normalized and summarized using the function three step of affyPLM R package (www.bioconductor.org) with default settings, background correction, quantile normalization and median polish probe summarization. ALMAC gene expression profiles from the PETACC-3 trial were processed as previously described (Popovici et al., 2011; Popovici et al., 2012). In each dataset, one probeset with the highest variability was selected as a representative of each EntrezGene ID. The variability for each probeset was estimated by robust linear regression (rlm function in R package MASS) as the robust scale estimate (RSE). This resulted in the following number of EntrezGene IDs: 21,758 in GEMM datasets, 14,926 in PETACC-3 dataset, 20,752 in GSE 14333 dataset and 11,237 in the cell line dataset. For all analyses with clinical data, an overlapping set of 13,265 EntrezGene IDs between the two clinical datasets (from ALMAC and Affymetrix platforms) was used. For signature development, mouse EntrezGene IDs were matched to their human homologs, reducing the number of EntrezGene IDs to 15,888 and intersected with 13,265 EntrezGene IDs of clinical datasets, leading to a final subset of 11,745 EntrezGene IDs.

Statistical analysis, clustering and classifier development

A multivariable linear additive model was built on a GEMM training set of 15,888 EntrezGene IDs to estimate mutation-allele-specific (*Apc*, *Kras*, *Braf*, *Tp53*) effects, with WT in all alleles as baseline. The genes that were assigned a statistically significant effect in a given mutation made up the mutation-specific gene list. Unsupervised hierarchical clustering with average linkage and Pearson correlation as a measure of similarity was used to cluster sets of the top 500 most variable EntrezGene IDs and then the top 500 most variable allele-specific genes and samples. For classifier construction, the final subset of 11,745 human homolog EntrezGene IDs was used.

The top 100 up- and downregulated genes from multivariable analysis specific for a given allele were used to define the allele-specific score, defined as a difference of average gene expression between up- and downregulated genes of the allele. The rule score >0 served as classifier defining allele-like group, except for the *KRAS* mutant subpopulation, where the median of the *KRAS*-like score was taken as threshold. Prior to application of the classifier and consequent survival analysis, the genes in the datasets were median-centered and normalized by inter-quartile range.

MSigDB analysis

Gene lists associated with each mutant allele (*Kras*, *Braf*, *Apc*, *Tp53*) generated from the multivariable analysis above ($P<0.01$ regulated for each allele) were uploaded to the MSigDB analysis tool [Broad Institute (http://www.broadinstitute.org/gsea/msigdb/index.jsp)]. Enrichment in MSigDB gene sets from all major canonical pathway collections were assessed and ranked by *P*-value. The top 10-20 MSigDB gene sets with the most significant enrichment for each allelic gene list were identified.

Comparison of GEMM *Kras* signature score and cell line sensitivity

GEMM *Kras* signature score classifier was applied to normalized, EntrezGene ID summarized cell line dataset (http://www.cancerrxgene.org). For this purpose, 66 upregulated and 74 downregulated EntrezGene IDs from the original GEMM *Kras* classifier that were found on the Affymetrix HG U133A platform were used to calculate the GEMM *Kras* score for each CRC cell line in this dataset. This score was then plotted with the corresponding IC_{50} values of drug response to the MEK inhibitors PD-0325901 and AZD6244 for each cell line, as reported in this dataset, and a linear model was fitted.

Independent confirmation of cell line sensitivity to MEK inhibitors

An independent validation of sensitivity to MEK inhibitors PD-0325901 and AZD6244 based on GEMM *Kras* signature score was performed by selecting representative cell lines with relatively high GEMM *Kras* signature scores (LS-1034, LS-513) and low signature scores (Colo320, SW948). Briefly, cell lines were seeded at 1000 cells/well in 96-well culture plates in growth medium with 10% FBS. Cells were incubated overnight and treated with DMSO (0.1% final) or serial diluted compound for 4 days. Cell viability was assessed adding Cell Titer Glo reagent (CTG, Promega, Madison, WI) and plates were incubated at room temperature for 30 minutes. Luminescence signals were read and IC_{50} values were calculated by plotting luminescence intensity to drug concentration in nonlinear curves using GraphPad Prism (GraphPad, La Jolla, CA).

Survival analysis

Outcome variables were overall survival (OS), relapse-free survival (RFS) and survival after relapse (SAR). Survival probabilities were estimated using the Kaplan-Meier method, and Cox proportional hazards model and Wald test were used to assess association of GEMM *Kras* signature with outcome variables. Cox proportional hazards model was used also for multivariable model. Survival times were cut at 84 months.

Gene expression data

All gene expression data can be found at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE50794.

Acknowledgements

We would like to thank the Pfizer postdoctoral program for their support.

Competing interests

P.J.B., P.J., T.X., P.A.R., K.E.H. and E.S.M. are either former or current employees of Pfizer.

Author contributions

E.S.M., O.J.S. and K.E.H. conceived the idea for the mouse models; E.S.M., O.J.S., K.E.H., E.C., J.R., M.J.S. and P.J.B. generated the mouse models; P.J.B., E.B., P.J., E.C., M.J.S., S.D., K.E.H., T.X., P.A.R., O.J.S., M.D., S.T. and E.S.M. characterized the mouse models; E.B. performed statistical and computational analyses; P.J.B., E.B. and E.S.M. wrote the manuscript; E.C., M.J.S., K.E.H., T.X., P.A.R., O.J.S., M.D., S.T. and E.S.M. reviewed and commented on the manuscript.

Funding

This work was supported by Pfizer Global Research and Development, the Swiss National Science Foundation (SNF; http://www.snf.ch) 320030_135421 (M.D.); Krebsforschung Schweiz (KFS; http://www.krebsforschung.ch/) 02697-08-2010 (M.D.); Fondation Medic (E.B. and M.D.) and Swiss National Center for Competence in Research Molecular Biology (http://www.nccr-oncology.ch/) (M.D.).

Supplementary material

Supplementary material available online at http://dmm.biologists.org/lookup/suppl/doi:10.1242/dmm.013904/-/DC1

References

Arnold, C. N., Goel, A., Blum, H. E. and Boland, C. R. (2005). Molecular pathogenesis of colorectal cancer: implications for molecular diagnosis. *Cancer* **104**, 2035-2047.

Betensky, R. A., Louis, D. N. and Cairncross, J. G. (2002). Influence of unrecognized molecular heterogeneity on randomized clinical trials. *J. Clin. Oncol.* **20**, 2495-2499.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Hodgson, J. G., Weinrich, S. et al. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **231**, 63-76.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337.

Clevers, H. (2006). Wnt/beta-catenin signaling in development and disease. *Cell* **127**, 469-480.

Coffee, E. M., Faber, A. C., Roper, J., Sinnamon, M. J., Goel, G., Keung, L., Wang, W. V., Vecchione, L., de Vriendt, V., Weinstein, B. J. et al. (2013). Concomitant BRAF and PI3K/mTOR blockade is required for effective treatment of BRAF(V600E) colorectal cancer. *Clin. Cancer Res.* **19**, 2688-2698.

de Bono, J. S. and Ashworth, A. (2010). Translating cancer research into targeted therapeutics. *Nature* **467**, 543-549.

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P., de Jong, J. H., de Boer, O. J., van Leersum, R., Bijlsma, M. F. et al. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.* **19**, 614-618.

Dry, J. R., Pavey, S., Pratilas, C. A., Harbron, C., Runswick, S., Hodgson, D., Chresta, C., McCormack, R., Byrne, N., Cockerill, M. et al. (2010). Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.* **70**, 2264-2273.

DuPage, M., Dooley, A. L. and Jacks, T. (2009). Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat. Protoc.* **4**, 1064-1072.

Engelman, J. A., Chen, L., Tan, X., Crosby, K., Guimaraes, A. R., Upadhyay, R., Maira, M., McNamara, K., Perera, S. A., Song, Y. et al. (2008). Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. *Nat. Med.* **14**, 1351-1356.

Frese, K. K. and Tuveson, D. A. (2007). Maximizing mouse cancer models. *Nat. Rev. Cancer* **7**, 654-658.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J. et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158.

Hegde, G. V., de la Cruz, C. C., Chiu, C., Alag, N., Schaefer, G., Crocker, L., Ross, S., Goldenberg, D., Merchant, M., Tien, J. et al. (2013). Blocking NRG1 and other ligand-mediated Her4 signaling enhances the magnitude and duration of the chemotherapeutic response of non-small cell lung cancer. *Sci. Transl. Med.* **5**, 171ra118.

Heyer, J., Kwong, L. N., Lowe, S. W. and Chin, L. (2010). Non-germline genetically engineered mouse models for translational cancer research. *Nat. Rev. Cancer* **10**, 470-480.

Hung, K. E., Maricevich, M. A., Richard, L. G., Chen, W. Y., Richardson, M. P., Kunin, A., Bronson, R. T., Mahmood, U. and Kucherlapati, R. (2010). Development of a mouse model for sporadic and metastatic colon tumors and its use in assessing drug treatment. *Proc. Natl. Acad. Sci. USA* **107**, 1565-1570.

Johnson, L., Mercer, K., Greenbaum, D., Bronson, R. T., Crowley, D., Tuveson, D. A. and Jacks, T. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. *Nature* **410**, 1111-1116.

Jonkers, J. and Berns, A. (2002). Conditional mouse models of sporadic cancer. *Nat. Rev. Cancer* **2**, 251-265.

Jorissen, R. N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L. A., Arango, D., Kruhøffer, M. et al. (2009). Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin. Cancer Res.* **15**, 7642-7651.

Kirk, R. (2011). Genetics: In colorectal cancer, not all KRAS mutations are created equal. *Nat Rev Clin Oncol* **8**, 1.

Kirsch, D. G., Dinulescu, D. M., Miller, J. B., Grimm, J., Santiago, P. M., Young, N. P., Nielsen, G. P., Quade, B. J., Chaber, C. J., Schultz, C. P. et al. (2007). A spatially and temporally restricted mouse model of soft tissue sarcoma. *Nat. Med.* **13**, 992-997.

Kuraguchi, M., Wang, X. P., Bronson, R. T., Rothenberg, R., Ohene-Baah, N. Y., Lund, J. J., Kucherlapati, M., Maas, R. L. and Kucherlapati, R. (2006). Adenomatous polyposis coli (APC) is required for normal development of skin and thymus. *PLoS Genet.* **2**, e146.

Lee, J. T. and Gu, W. (2010). The multiple levels of regulation by p53 ubiquitination. *Cell Death Differ.* **17**, 86-92.

Levine, A. J. and Puzio-Kuter, A. M. (2010). The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* **330**, 1340-1344.

Martin, E. S., Belmont, P. J., Sinnamon, M. J., Richard, L. G., Yuan, J., Coffee, E. M., Roper, J., Lee, L., Heidari, P., Lunt, S. Y. et al. (2013). Development of a colon cancer GEMM-derived orthotopic transplant model for drug discovery and validation. *Clin. Cancer Res.* **19**, 2929-2940.

Politi, K. and Pao, W. (2011). How genetically engineered mouse tumor models provide insights into human cancers. *J. Clin. Oncol.* **29**, 2273-2281.

Pollock, C. B., Shirasawa, S., Sasazuki, T., Kolch, W. and Dhillon, A. S. (2005). Oncogenic K-RAS is required to maintain changes in cytoskeletal organization, adhesion, and motility in colon cancer cells. *Cancer Res.* **65**, 1244-1250.

Popovici, V., Budinská, E. and Delorenzi, M. (2011). Rgtsp: a generalized top scoring pairs package for class prediction. *Bioinformatics* **27**, 1729-1730.

Popovici, V., Budinska, E., Tejpar, S., Weinrich, S., Estrella, H., Hodgson, G., Van Cutsem, E., Xie, T., Bosman, F. T., Roth, A. D. et al. (2012). Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J. Clin. Oncol.* **30**, 1288-1295.

Racker, E., Resnick, R. J. and Feldman, R. (1985). Glycolysis and methylaminoisobutyrate uptake in rat-1 cells transfected with ras or myc oncogenes. *Proc. Natl. Acad. Sci. USA* **82**, 3535-3538.

Rajalingam, K., Schreck, R., Rapp, U. R. and Albert, S. (2007). Ras oncogenes and their downstream targets. *Biochim. Biophys. Acta* **1773**, 1177-1195.

Roper, J. and Hung, K. E. (2012). Priceless GEMMs: genetically engineered mouse models for colorectal cancer drug development. *Trends Pharmacol. Sci.* **33**, 449-455.

Roth, A. D., Tejpar, S., Delorenzi, M., Yan, P., Fiocca, R., Klingbiel, D., Dietrich, D., Biesmans, B., Bodoky, G., Barone, C. et al. (2010). Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *J. Clin. Oncol.* **28**, 466-474.

Rudrapatna, V. A., Cagan, R. L. and Das, T. K. (2012). Drosophila cancer models. *Dev. Dyn.* **241**, 107-118.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C., Lannon, W. A., Grotzinger, C., Del Rio, M. et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**, 619-625.

Sumimoto, H., Imabayashi, F., Iwata, T. and Kawakami, Y. (2006). The BRAF-MAPK signaling pathway is essential for cancer-immune evasion in human melanoma cells. *J. Exp. Med.* **203**, 1651-1656.

Tuveson, D. and Hanahan, D. (2011). Translational medicine: Cancer lessons from mice to humans. *Nature* **471**, 316-317.

Tuveson, D. A. and Jacks, T. (2002). Technologically advanced cancer modeling in mice. *Curr. Opin. Genet. Dev.* **12**, 105-110.

Van Cutsem, E., Labianca, R., Bodoky, G., Barone, C., Aranda, E., Nordlinger, B., Topham, C., Tabernero, J., André, T., Sobrero, A. F. et al. (2009). Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J. Clin. Oncol.* **27**, 3117-3125.

Vidal, M. and Cagan, R. L. (2006). Drosophila models for cancer research. *Curr. Opin. Genet. Dev.* **16**, 10-16.

Yun, J., Rago, C., Cheong, I., Pagliarini, R., Angenendt, P., Rajagopalan, H., Schmidt, K., Willson, J. K., Markowitz, S., Zhou, S. et al. (2009). Glucose deprivation contributes to the development of KRAS pathway mutations in tumor cells. *Science* **325**, 1555-1559.

[***11***] Byrne AT, Alférez DG, Amant F, Annibali D, Arribas J, Biankin AV, Bruna A, **Budinská E**, Caldas C, Chang DK, Clarke RB, Clevers H, Coukos G, Dangles-Marie V, Eckhardt SG, Gonzalez-Suarez E, Hermans E, Hidalgo M, Jarzabek MA, de Jong S, Jonkers J, Kemper K, Lanfrancone L, Mælandsmo GM, Marangoni E, Marine JC, Medico E, Norum JH, Palmer HG, Peeper DS, Pelicci PG, Piris-Gimenez A, Roman-Roman S, Rueda OM, Seoane J, Serra V, Soucek L, Vanhecke D, Villanueva A, Vinolo E, Bertotti A, Trusolino L. Interrogating open issues in cancer precision medicine with patient-derived xenografts. Nat Rev Cancer. 2017 Apr;17(4):254-268. doi: 10.1038/nrc.2016.140. Epub 2017 Jan 20. Erratum in: Nat Rev Cancer. 2017 Sep 15;17(10):632. doi: 10.1038/nrc.2017.85. PMID: 28104906.

OPINION

Interrogating open issues in cancer precision medicine with patient-derived xenografts

Annette T. Byrne, Denis G. Alférez, Frédéric Amant, Daniela Annibali, Joaquín Arribas, Andrew V. Biankin, Alejandra Bruna, Eva Budinská, Carlos Caldas, David K. Chang, Robert B. Clarke, Hans Clevers, George Coukos, Virginie Dangles-Marie, S. Gail Eckhardt, Eva Gonzalez-Suarez, Els Hermans, Manuel Hidalgo, Monika A. Jarzabek, Steven de Jong, Jos Jonkers, Kristel Kemper, Luisa Lanfrancone, Gunhild Mari Mælandsmo, Elisabetta Marangoni, Jean-Christophe Marine, Enzo Medico, Jens Henrik Norum, Héctor G. Palmer, Daniel S. Peeper, Pier Giuseppe Pelicci, Alejandro Piris-Gimenez, Sergio Roman-Roman, Oscar M. Rueda, Joan Seoane, Violeta Serra, Laura Soucek, Dominique Vanhecke, Alberto Villanueva, Emilie Vinolo, Andrea Bertotti and Livio Trusolino

Abstract | Patient-derived xenografts (PDXs) have emerged as an important platform to elucidate new treatments and biomarkers in oncology. PDX models are used to address clinically relevant questions, including the contribution of tumour heterogeneity to therapeutic responsiveness, the patterns of cancer evolutionary dynamics during tumour progression and under drug pressure, and the mechanisms of resistance to treatment. The ability of PDX models to predict clinical outcomes is being improved through mouse humanization strategies and the implementation of co-clinical trials, within which patients and PDXs reciprocally inform therapeutic decisions. This Opinion article discusses aspects of PDX modelling that are relevant to these questions and highlights the merits of shared PDX resources to advance cancer medicine from the perspective of EurOPDX, an international initiative devoted to PDX-based research.

Response to anticancer therapies varies owing to the substantial molecular heterogeneity of human tumours and to poorly defined mechanisms of drug efficacy and resistance[1]. Immortalized cancer cell lines, either cultured *in vitro* or grown as xenografts, cannot interrogate the complexity of human tumours and only provide determinate insights into human disease, as they are limited in number and diversity, and have been cultured on plastic over decades[2]. This disconnection in scale and biological accuracy contributes considerably to attrition in drug development[3–5].

Surgically derived clinical tumour samples that are implanted in mice (known as patient-derived xenografts (PDXs)) are expected to better inform therapeutic development strategies. As intact tissue — in which the tumour architecture and the relative proportion of cancer cells and stromal cells are both maintained — is directly implanted into recipient animals, the alignment with human disease is enhanced. More importantly, PDXs retain the idiosyncratic characteristics of different tumours from different patients; hence, they can effectively recapitulate the intra-tumour and inter-tumour heterogeneity that typifies human cancer[6–9].

Exhaustive information on the key characteristics and the practical applications of PDXs can be found in recent reviews[10–13]. In this Opinion article, we discuss basic methodological concepts, as well as challenges and opportunities in developing ‘next-generation’ models to improve the reach of PDXs as preclinical tools for *in vivo* studies (TABLE 1). We also elaborate on the merits of PDXs for exploring the intrinsic heterogeneity and subclonal genetic evolution of individual tumours, and discuss how this may influence therapeutic resistance. Finally, we examine the utility of PDXs in navigating complex variables in clinical decision-making, such as the discovery of predictive and prognostic biomarkers, and the categorization of genotype–drug response correlations in high-throughput formats. Being primarily co-authored by leading members of the EurOPDX Consortium (see Further information), we provide a perspective on the value of PDX models as an important resource for the international cancer research community towards the realization of a precision medicine paradigm (BOX 1; TABLE 2).

Modelling cancer phenotypes

Interrogating intra-tumour heterogeneity and evolutionary dynamics. Cancer is increasingly being recognized as an ecosystem of cells that constantly evolves following Darwinian laws. Owing to cancer cell intrinsic mutability, an incipient tumour clone gives rise to a progeny of genetically heterogeneous subclones, some of which will thrive while others shrink, depending on their ability to cope with environmental selection pressures[14]. This is of particular relevance for cancer treatment, as most patients will eventually succumb to the disease owing to the appearance of resistant tumour subclones. Despite the considerable clinical impact of tumour heterogeneity[15], little is known about how it affects response to cancer therapy and how it may change during treatment at both the genomic and the phenotypic levels[16–20]. These issues highlight the need for preclinical models that capture the heterogeneous nature of human cancers and their ongoing evolution.

Table 1 | **Modelling cancer phenotypes with PDX models**

PDX model	Open clinical question	Advantages	Challenges
Primary tumour specimens implanted s.c.	• Interrogation of primary or acquired resistance mechanisms • Discovery of prognostic and predictive biomarkers • Drug response • Identification of targetable molecular alterations • Characterization of intra-tumour clonal evolution	• Intact primary tumour tissue that maintains tumour architecture • Captures clonal diversity • Easy to measure tumour responses • Intravital tumour imaging	• Lack of proper anatomical niche • Not all grades of tumour engraft s.c. Generally, higher grade, more aggressive tumours engraft more easily
Primary tumour specimens implanted orthotopically (PDOX)	• Mechanisms of metastasis • Study site-specific dependence of therapy • Monitoring the effects of adjuvant therapy on occult metastasis • Stromal contribution to response	• Intact primary tumour tissue that maintains primary tumour architecture • Local growth of primary tumour in proper anatomical context • Spontaneous distant metastases from primary tumour • Presence of primary and metastatic tumour niche • Recapitulates the entire metastatic process from the appropriate anatomical site • Ability to mimic clinical scenarios, for example, surgical removal of primary tumour or adjuvant therapy	• Access to imaging technologies to visualize tumour in longitudinal studies • Microsurgical skills • Large collections and high-throughput screens difficult to implement
Metastatic tumour specimens implanted s.c.	• Interrogation of primary or acquired resistance mechanisms • Discovery of prognostic and predictive biomarkers • Drug response • Identification of targetable molecular alterations • Characterization of intra-tumour clonal evolution	Intact metastatic tumour tissue that maintains tumour architecture	Lack of tumour metastatic niche
Metastatic tumour specimens implanted orthotopically at the metastatic site	• Mechanisms of metastasis • Drug resistance • Genetic and cellular mechanisms of tumour growth • Drug response in the setting of metastatic disease • Stromal contribution to response	Intact metastatic tumour tissue that maintains tumour architecture	• Access to imaging technologies to visualize tumour in longitudinal studies • Microsurgical skills • Large collections and high-throughput screens difficult to implement
PDX models of MRD	• Drug resistance • Discovery of prognostic and predictive biomarkers • Biological and pharmacological studies • Identification of targetable molecular alterations	• Studies can help us to understand the molecular bases of and optimize therapies for MRD • Higher tumour take rate when compared with untreated cancers • Enables the study of clonal evolution and cancer stem cell behaviour	PDXs are never therapy naive
Clinical trial-associated xenografts (CTAXs)	• Discovery of prognostic and predictive biomarkers • Drug resistance • Drug response • Identification of targetable molecular alterations • Mechanisms of metastasis	• Possibility of establishing xenografts at different clinical stages during patient tumour progression • Permits the parallel testing of novel drug combinations	• Limited quantity and quality of tissue • Limited number of successfully generated PDXs • A PDX derived from a single biopsy sample may not represent the patient's tumour
CTC-derived PDX models	• Molecular tumour heterogeneity • Discovery of prognostic and predictive biomarkers • Study of the genetic evolution of the tumour • Identification of targetable molecular alterations	• Minimally invasive sampling • Ability to monitor cancer burden and drug susceptibility in metastatic and late-stage settings • Recapitulates donor patient's response to treatment • Facilitates investigation of the biology of otherwise inaccessible tumour specimens	• Low concentration in peripheral blood of patients with different solid tumours • Access to technologies to isolate all CTCs (both epithelial and mesenchymal) • Technically challenging
Humanized PDX models	Investigation of immune therapeutics	Recapitulates human immune system in mice	• Requires lengthy mouse humanization procedures • Hurdles to achieve complete human immune system reconstitution • See Supplementary information S1 (table) for further details

CTC, circulating tumour cell; MRD, minimal residual disease; PDX, patient-derived xenograft; PDOX, patient-derived orthotopic xenograft; s.c., subcutaneously.

Box 1 | The EurOPDX Consortium and other related initiatives

EurOPDX (see Further information; established in 2013) is a collaborative network of 16 European academic institutions with expertise in basic, preclinical, translational and clinical oncology. Participating laboratories are affiliated with comprehensive cancer centres within which preclinical experimentation is closely associated with clinical activities. This allows for the efficient sharing of patient specimens — together with fully annotated clinical information — and facilitates the collection of tumours with unique characteristics (for example, rare types, exceptional responders and therapy-refractory cases). Currently sustained by membership fees organized by a consortium agreement, EurOPDX aims to obtain competitive infrastructural funding to further implement collaborative research projects and to formalize external access procedures to models. The consortium agreement also sets forth general rules for confidentiality and intellectual property issues to regulate activities among EurOPDX members (co-ownership of results) and between EurOPDX and potential partners, including other patient-derived xenograft (PDX) consortia and industry.

Main objectives

- To create a uniquely extensive collection of characterized PDX models. The collection consists of more than 1,500 subcutaneous and orthotopic models from more than 30 different pathologies (see TABLE 2). The models and their molecular annotation are currently being made publicly available through the cBioPortal, and are accessible for collaboration upon signature of a material transfer agreement. Systematic derivation of primary cultures and organoids for *in vitro* studies is planned.
- To provide a platform for population-scale studies to discover low-prevalence genetic alterations with clinically actionable potential; to explore mechanisms of therapeutic resistance in molecularly defined tumour subtypes; and to develop predictive biomarkers for personalized cancer treatment.
- To harmonize working practices. This entails several aspects: first, standardization of biobanking procedures, including systematic assessment of genetic identity by single nucleotide polymorphism (SNP) DNA fingerprinting. Second, the implementation of common rules for PDX expansion and archiving; discussions are ongoing to limit PDX propagation to a maximum of five passages, but exceptions will be considered for tumour types known to deteriorate after freezing–thawing steps and for models characterized by very indolent growth, for which expansions up to five passages would take exceedingly long. Third, optimization of inter-laboratory reproducibility through proof-of-concept studies by which models from the same source are tested independently. And finally, the definition of a set of minimal information to be linked to each PDX.

Other major PDX initiatives

- US National Cancer Institute (NCI) repository of patient-derived models (see Further information).
- US Pediatric Preclinical Testing Consortium (PPTC; see Further information), a US National Cancer Institute (NCI)-centralized and NCI-funded collection for *in vivo* testing of paediatric anticancer drug candidates.
- Children's Oncology Group (COG) cell culture and xenograft repository (see Further information), a COG-based resource that provides validated cell lines and PDXs from paediatric cancers.
- Public Repository of Xenografts (PRoXe; see Further information), an open-source repository of leukaemia and lymphoma PDXs[165]. Many of the models are being licensed to the Jackson Laboratories for industry-scale purposes, including distribution on a fee-for-service basis.
- Novartis Institutes for Biomedical Research PDX Encyclopedia (NIBR PDXE), an industry-led initiative that includes approximately 1,000 models[9]. Clinical, pathological and PDX-level data from this collection are currently being incorporated into PRoXe[165].

For example, breast cancer is a constellation of at least 10 different genomic subtypes, each with distinct drivers and variable intra-tumour heterogeneity[15,21,22]. Recent evidence has suggested that each breast cancer comprises multiple tumour cell populations with distinct evolutionary trajectories that are likely to be affected by treatment pressure[23–25]. Genomic evolution between primary and recurrent tumours also occurs[24–28]. Such intra-tumour and inter-tumour variability affects therapeutic responses, and hence needs to be considered in the preclinical and clinical settings. Although some engraftment-associated selection has been documented[24,29], PDX models of breast cancer seemingly preserve most of the genomic clonal architecture of the original patient sample and also seem to resemble patient counterparts at the transcriptomic, epigenomic and histological levels, as well as in terms of shared signalling pathways[8,30–32]. Notably, the majority of tumour subclones that change upon engraftment do not include known breast cancer oncogenic drivers[29]. This suggests that, although engraftment pressure is observed, it is evolutionarily neutral, as it does not affect intra-tumour heterogeneity when considering the clonal representation of relevant genes. These features probably underpin the successful use of breast cancer PDXs to predict clinical drug responses[9] and mechanisms of acquired resistance[33,34].

As discussed below, an advantage of PDX models is that they can be generated with a limited amount of material; for example, using fine-needle biopsies (TABLE 1). However, these methods may be confounding when the studied tumour type is particularly heterogeneous (such as melanoma). For example, within one tumour or metastasis, multiple melanoma subclones can exist, each harbouring different genetic and/or epigenetic alterations[35–37]. Simply taking a single biopsy sample can result in a PDX that does not represent the heterogeneity of the patient's tumour[35,38]. Notably, regional genetic variability can be exacerbated by PDX serial propagation, producing divergent responses in tumour measurements within a single cohort of treated mice[32,39]. Methods to overcome this limitation include good, standardized preclinical designs (those with adequate statistical power and proper randomization), as well as the mixing of heterogeneous tumour masses before implantation, such as through the use of single-cell suspension injections or rough tumour homogenates[24].

The direct derivation of PDXs from circulating tumour cells (CTCs) may represent another tool to further interrogate tumour heterogeneity. The numbers of cancer cells shed by tumours into the bloodstream may be exceedingly low, and the biological and clinical relevance of CTCs in sustaining malignant disease has been questioned[40]. However, as CTCs are shed by tumours on a stochastic rather than a deterministic basis[41], they are expected to better recapitulate the distribution of different subclonal tumour populations (TABLE 1).

Intra-tumoural heterogeneity may also be non-genetic and intrinsic to the hierarchical organization of some tumours, in which a small subpopulation of cancer stem cells (CSCs) may be responsible for long-term tumorigenicity[42–45]. CSCs are thought to be chemoresistant and the main cause of recurrence and distant metastasis[46–48]. Much of the supporting evidence originates from PDX models that were directly derived from various clinical samples, including CTCs, ascites fluid and pleural effusion cells, and surgical biopsy samples[49–53]. PDX models have provided evidence of CSC colonization in metastatic sites and have also highlighted the role and importance of the surrounding tumour stroma, a niche that is known to influence CSC behaviour by cell-to-cell contacts and through the secretion of pro-tumorigenic ligands and cytokines[8,51,54]. An ongoing debate exists as to whether CSCs recapitulate the full characteristics of stem cells (that is, they are undifferentiated cells with limitless replicative potential, which partly self-perpetuate to maintain a tumorigenic reservoir and which partly differentiate to give rise to a diverse progeny of non-tumorigenic cells) or simply identify

Table 2 | **Facts and figures about the EurOPDX collection***

Tumour type or organ	Subtype	Primary tumour or metastasis	Total number of established models	Average engraftment rate: treatment naive and adjuvant samples (%)		Engraftment rates: neoadjuvant samples (if relevant) (%)	
				Subcutaneous	Orthotopic	Subcutaneous	Orthotopic
CRC	All subtypes included	Primary	291	52–75	80	NA	NA
		Liver metastasis	444	73–91	90	84	NA
Pancreas (PDAC)	All subtypes included	Primary	211	54–71	70	NA	NA
		Liver metastasis	24	60–100	90	NA	NA
Breast	ER$^+$ (including ER$^+$HER2$^+$)	Primary	24	4–7	7	20	NA
		Metastasis	20	25–49	33–47	NA	NA
	TNBC	Primary	78	30–34	60–86	72	86
		Metastasis	26	60	50–66	NA	NA
	HER2$^+$ only	Primary	16	26	NA	13	NA
		Metastasis	5	NA	33	NA	NA
Skin melanoma	All subtypes included	Primary	8	67–90	29	NA	NA
		Metastasis (cutaneous, liver and lung)	161	72–90	83–85	NA	NA
Ovary	All subtypes included	Primary	123	40–85[‡]	68	62[‡]	NA
		Metastasis	19	47–85[‡]	80	NA	NA
Gastric	All subtypes included	Primary	87	41–50	70	34	NA
Endometrial	All subtypes included	Primary	67	43–55	74	NA	NA
		Metastasis	10	10–60[§]	95	NA	NA
Lung	NSCLC	Primary and metastasis	59	50–70 (primary)	52	NA	NA
	SCLC	Primary and metastasis	12	50	75	Not applicable	Not applicable
HNSCC	All subtypes included	Primary	50	45	65	NA	NA
		Metastasis	13	83	NA	NA	NA
Glioblastoma	All subtypes included	Primary	52	Not applicable	95–100	Not applicable	NA
Uveal melanoma	All subtypes included	Primary	12	32	NA	Not applicable	Not applicable
		Liver metastasis	14	65	NA	Not applicable	Not applicable
Testicular	All subtypes included	Primary and metastasis (lymph node, lung and brain)	18	NA	35	NA	NA
Uterine sarcoma	High grade	Primary	3	75	NA	Not applicable	Not applicable
		Metastasis	9	100	NA	Not applicable	Not applicable
Renal	All subtypes included	Primary	8	30	NA	NA	NA

CRC, colorectal cancer; ER, oestrogen receptor; HNSCC; head and neck squamous cell carcinoma; NA, not available; NSCLC, non-small-cell lung cancer; PDAC, pancreatic ductal adenocarcinoma; SCLC, small-cell lung cancer; TNBC, triple-negative breast cancer. *The data presented represent the range of implantation rates obtained across EurOPDX partner laboratories as of October 2016. [‡]Highest take rates obtained with the high-grade serous ovarian cancer subtype. [§]Take rates of 10–15% for abdominal, pelvic lymph node and peritoneal metastases, 60% for vaginal metastases.

a more robust or proliferative population of 'tumour-initiating' cells selected by engraftment. To address this quandary, it will be important to compare the results of side-by-side fate-mapping experiments and transplantation assays to analyse whether the cells endowed with tumorigenic potential after transplantation also exhibit other typical stem-like properties, such as the ability to self-renew, asymmetric cell division and differentiation potential[55].

PDX models of treatment-resistant disease. There are primarily two ways in which PDX models can be used to interrogate primary and acquired resistance. One strategy is to derive models from patients' samples before the initiation of therapy and again at the time of treatment resistance. Alternatively, models can be developed from pretreatment tumour samples, and resistance can be recapitulated in the PDX upon iterative cycles of exposure to the drug, as previously observed in genetically engineered mouse (GEM) models[56]. Using cycles of drug exposure in pretreatment PDX models, paired analysis of PDX models of cisplatin-sensitive and cisplatin-resistant testicular germ cell cancer (TGCC) proposed potential alternatives for the treatment of cisplatin-refractory TGCC, including anti-angiogenic therapy[57] and the blockade of the platelet-derived growth factor receptor-β (PDGFRβ)–AKT pathway[58].

PDX models have also proved useful in identifying mechanisms of resistance to targeted therapies in oestrogen receptor (ER)-positive breast cancer. The analysis of four hormone-resistant PDX tumours, which were obtained from two ER-positive breast cancer PDX models by continuous treatment with tamoxifen or by oophorectomy-mediated hormone depletion, revealed that hormone resistance was associated with various forms of deregulated ER-mediated gene transcription[33]. Taking a similar approach, PDX models of ER-positive breast cancer have been used to investigate jagged 1 (JAG1)–NOTCH4 signalling as a means for attenuating sensitivity to hormonal therapy[59] and to identify mechanisms of acquired resistance to cyclin-dependent kinase 4 (CDK4) and CDK6 blockade[60].

Patients with advanced cancer who acquire resistance to several lines of treatment mostly present with multiple metastatic lesions that are not amenable to resection, and may harbour different resistance pathways. Generating PDX models that recapitulate such complex scenarios of therapy-resistant metastatic tumours has become feasible for several tumours (TABLE 1). For example, the analysis of biopsy specimens and corresponding PDXs from different drug-resistant metastases in patients with melanoma who had been treated with a BRAF inhibitor resulted in the identification of multiple resistance mechanisms both within individual lesions and among separate samples from the same patient[35,38]. The resistance mechanisms identified in PDXs were also found in the original patient samples[35], and clinically resistant tumours were also treatment-refractory when grown as PDXs[38]. These studies provide proof of principle for the heterogeneous nature of acquired resistance in individual patients with melanoma and further attest to the ability of PDX models to predict clinical outcomes. Similar results have been observed in lung adenocarcinomas[61].

Although PDXs generally retain drug-sensitivity profiles that are similar to those of the corresponding patient tumour[30,38,62,63], PDX models derived from treatment-resistant tumours can become sensitive again upon xenografting, owing to the effect of the so-called 'drug holiday' in which treatment is discontinued after tumour implantation to facilitate engraftment. Some resistance mechanisms are thus reversible in the absence of drug, as shown for melanoma[64,65] and lung adenocarcinoma[66]. This suggests that treatment-resistant PDXs should be exposed to continuous treatment immediately after implantation, although this is a cost- and labour-intensive approach. However, uninterrupted therapy might also result in the further selection of a subpopulation of tumour cells, resulting in a loss of intra-tumour heterogeneity and genetic variation in the PDX tumours compared with the original tumours.

In response to the need for more sophisticated models, several groups (for example, see REF. 67) have developed protocols and networks to generate clinical trial-associated xenografts (CTAXs) (TABLE 1). These advanced PDX models are currently being derived from image-guided biopsy samples taken at different time points during disease progression and following new lines of treatment in the context of clinical trials. Such models will be extremely valuable in evaluating how the molecular evolution of advanced tumours is associated with innate or acquired drug resistance, and will be important for studying the tumour heterogeneity and clonal selection that results from drug treatment. In principle, CTAXs may also serve as personalized cancer models to test drug combinations that aim to overcome acquired resistance, generating information that could be transferred back to the donor patient for therapeutic decisions (see below). However, this opportunity might be hindered by limitations such as the low engraftment success rates for some tumour types and the disconnection between the time needed for PDX expansion and treatment (which can be long, especially for tumours with indolent growth in mice) and the rapidity of disease progression in patients.

Finally, PDXs that are established from tumours resistant to conventional therapies delivered in the neoadjuvant setting are of special interest (TABLE 1). In triple-negative breast cancer, the establishment and molecular profiling of PDXs from residual cancer cells that persist after neoadjuvant treatment (minimal residual disease (MRD)) may lead to the identification of targetable molecular alterations in the chemotherapy-resistant component of the tumour, which may mirror micro-metastases that are destined to clinically recur [68]. Despite often being limited in size due to prior exposure to cytotoxic therapy, triple-negative breast tumours from patients treated in the neoadjuvant setting engraft much more efficiently than do treatment-naive tumours (72% and 34%, respectively) (TABLE 2). Given the high engraftment efficiency and rapid growth of PDXs from drug-tolerant MRD tissues, at least in the case of breast cancer, these models represent an unprecedented opportunity to identify genomic alterations and associated targeted therapies before tumour recurrence in patients.

Next-generation PDX models

Humanized PDX models to evaluate cancer immunotherapies. The importance of the immune system in tumour progression and treatment highlights the need for PDX models to facilitate the preclinical assessment of cancer immune therapies[69]. However, to avoid immune rejection of xenotransplants by the host, PDX models are primarily generated by transplanting tumour fragments into immunodeficient mice. The absence of many components of the immune system in these mice, and the loss of endogenous human immune cells upon propagation of the human tumour tissue over multiple passages[70,71], limit the utility of such models to explore the role of the immune system in tumour progression and to test novel immune-based therapies[72].

Humanized mice (also known as human haemato-lymphoid chimeric mice and human immune system (HIS) models)

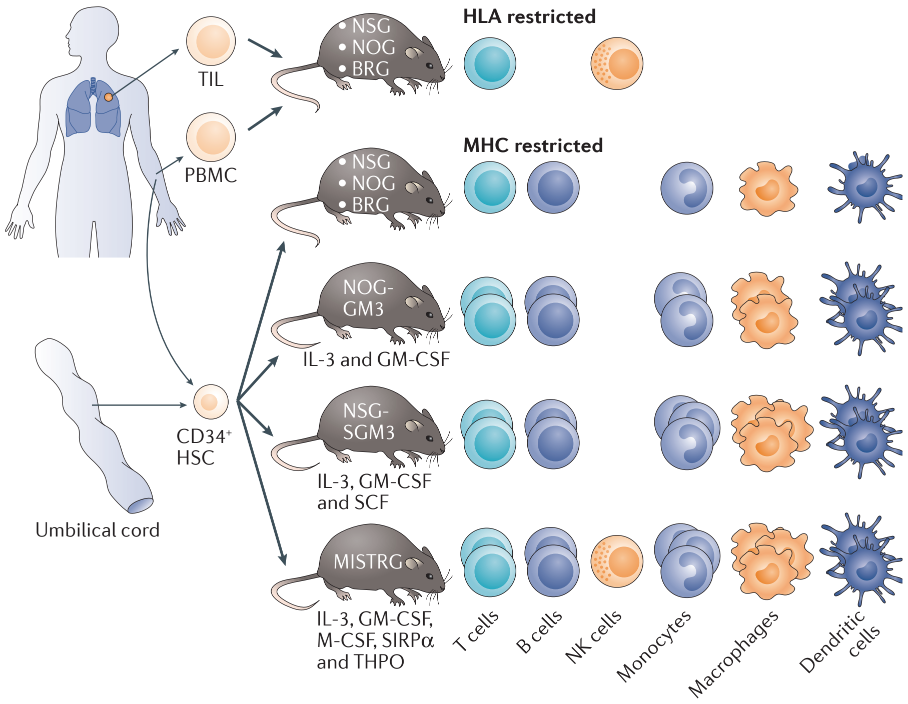


Figure 1 | **Strategies to generate humanized PDXs.** Sources of immune cells include tumour-infiltrating lymphocytes (TILs), peripheral blood mononuclear cells (PBMCs) and CD34-positive haematopoietic stem cells (HSCs); HSCs may be purified from mobilized adult peripheral blood, bone marrow or umbilical cord blood. Engrafted TILs or PBMCs generate mainly circulating human leukocyte antigen (HLA)-restricted T cells and natural killer (NK) cells (top row). This system is characterized by a vigorous graft-versus-host reaction that narrows the experimental window to approximately 2–5 weeks. Despite this limitation, the system is useful for certain analyses, such as monitoring the recruitment of T lymphocytes to tumours by therapeutic antibodies[170]. Fully humanized systems (bottom four rows) use severely immunodeficient mouse strains such as NOG (NOD-Cg-*Prkdc*^{scid} *Il2rg*^{tm1Sug}/JicTac)[171], NSG (NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ)[172] and BRG (C.Cg-Rag1^{tm1Mom} IL2rg^{tm1Wjl}/SzJ)[173,174]. Mice with a NOD (non-obese diabetic) background have functionally deficient NK cells. SCID (severe combined immunodeficiency) is a loss-of-function mutation that affects DNA-dependent protein kinase (DNA-PK), a DNA repair enzyme involved in V(D)J recombination during T cell and B cell development. As a consequence, SCID mice have reduced levels of T cells and B cells. Inactivation of the interleukin-2 (IL-2) receptor γ-chain leads to impaired T cell and B cell development and prevents the generation of NK cells. Recombination-activating gene 1 (RAG1) is necessary for V(D) J recombination; thus, RAG1 inactivating mutations affect T cell and B cell development. All these different strains show subtle differences to support the engraftment of functional human immune cells[173]. Injection of human CD34-positive HSCs into these mice leads to the generation of major histocompatibility complex (MHC)-restricted T cells and B cells, as well as to limited amounts of monocytes, macrophages, neutrophils and dendritic cells. In addition, these mouse strains have been improved by genetic modifications for the production of a variety of human cytokines that stimulate the differentiation of additional haematopoietic lineages. For example, strains such as NOG-GM3 (which expresses human IL-3 and granulocyte–macrophage colony-stimulating factor (GM-CSF; also known as CSF2)[175], NSG-SGM3 (which expresses human IL-3, GM-CSF and SCF (also known as KIT ligand))[176] and MISTRG (which expresses IL-3, GM-CSF, macrophage CSF (M-CSF; also known as CSF1), signal regulatory protein-α (SIRPα) and thrombopoietin (THPO))[177] produce increased numbers of human myeloid and mast cells, regulatory T cells and NK cells (see Supplementary information S1 (table)). PDXs, patient-derived xenografts.

are immunocompromised mice in which selected immune components have been introduced to generate a competent human immune system with different degrees of immune reconstitution. One methodology for the generation of humanized mice involves the transplantation of total peripheral blood from healthy human donors or patients (peripheral blood lymphocyte (PBL) models) or, in particular applications, the infusion of tumour-infiltrating lymphocytes (TILs) (FIG. 1). Although these procedures are known to cause severe graft-versus-host disease (GvHD) beginning 2–5 weeks after injection[73,74], seriously limiting the useful investigative time window of these models and the translational value of these studies[75], PBL and TIL mice can be used for cost-effective short-term testing of novel immune therapeutics and for assessing short-term adverse effects.

Alternatively, HIS mice can be generated through the transplantation of CD34-positive human haematopoietic stem cells (HSCs) or precursor cells isolated from umbilical cord blood, bone marrow and peripheral blood, either alone or in combination with additional human immune tissues (bone ossicles or human thymic tissue)[76] into immunodeficient mice (FIG. 1). Compared with PBL- and TIL-derived models, transplantation with HSCs results in a more complete haematopoietic reconstitution, as HSCs give rise to various lineages of human blood cells throughout the life of the animal. Methods for transplantation depend on the source of HSCs, the co-transplantation of immune tissues, the mouse strain and the age of the recipient mice[75–78]. In order to avoid the immune reactions caused by human leukocyte antigen (HLA) mismatch, the ideal source of HSCs is the same patient from whom the PDX has been established. However, isolating HSCs from cancer patients may prove daunting: on the one hand, bone marrow biopsies are difficult in debilitated individuals; on the other hand, growth factor-stimulated bone marrow mobilization for HSC collection from peripheral blood might foster tumour progression[79]. Moreover, even when applicable, the low yield of HSCs obtainable from cancer patients severely limits the number of mice than can be humanized. An attractive alternative is the *in vitro* expansion of HSCs[80], although this procedure could introduce biological perturbations that affect stemness and differentiation potential.

Whereas various strains of immunodeficient mice are used to transplant solid tumour tissue, not all of these strains are suitable for generating HIS models. The survival of human immune cells is highly dependent on the compatibility of the 'do-not-eat-me' signals (CD47–signal-regulatory protein-α (SIRPα)) on phagocytes in the host[81]. The most commonly used mice to generate compatible HIS models are those derived from the non-obese-diabetic (NOD)-severe combined immune deficiency (SCID)-interleukin-2 receptor common γ-chain (IL2-Rγ)-deficient (NSG; also known as NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjll}/SzJ) strain and the NOD/Sci-SCID/IL-2Rγ strain (NOG; also known as NOD-Cg-*Prkdc*^{scid} *Il2rg*^{tm1Sug}/JicTac). Substantial efforts are thus being made to develop novel GEM strains that not only express human-specific do-not-eat-me signals but also express human-specific cytokines or HLAs. These mouse strains differ upon transplantation in durability and quality of engraftment of the human immune system[78]. Some key examples of how humanized models are currently evolving to support PDX transplantation towards application in the immune-oncology space are presented as online supplementary information (see Supplementary information S1 (table)).

Modelling metastatic disease. Subcutaneous transplantation usually fails to reproduce the organ-specific tropism of distant metastases that is observed in patients. Therefore, models of metastatic disease are typically generated through orthotopic procedures. These include the transplantation of fragments of the primary tumour into the same location in the mouse, which is usually followed by the development of spontaneous metastases, or the direct transfer of metastatic lesions into the same organ in the host (TABLE 1). Patient-derived orthotopic xenografts (PDOXs; also known as orthoxenografts) of primary tumours can reproducibly lead to local invasive growth and metastases, often identical to those observed in the patient[82–84]. PDOX models for most cancer indications have typically been developed from surgical specimens. More recently, however, they have been successfully derived from biopsy samples, despite the limited quantity and quality of tissue available[85].

Advantages of orthotopic models include the ability to investigate tumour–host interactions at the relevant site of primary and secondary tumour growth, the development of patient-like metastases, the ability to interrogate site-specific dependence of therapy, and the potential to conduct clinically relevant studies, such as monitoring the effects of adjuvant therapy on occult metastases (TABLE 1). Nevertheless, orthotopic models remain relatively rare, probably owing to the non-trivial microsurgical procedures that are required for organ-specific transplantation. Furthermore, the incorporation of clinically relevant imaging modalities and appropriate *in vivo* imaging probes is necessary to visualize tumour orthotopic implants and metastatic progression in deep tissues and to ensure timely therapeutic intervention when animals develop disease symptoms[86].

PDOX models of breast cancer are particularly amenable for modelling metastasis. They primarily rely on mammary fat pad injection of primary tumour samples, which successfully recapitulates the entire metastatic process from the appropriate primary anatomical site[8,87]. PDOX models of brain metastases and primary brain tumours are challenging. To prevent the default seeding of intravenously injected tumour cells in the lung and to ensure colonization of the central nervous system, intra-cardiac left ventricular inoculation of tumour cells is required[88]. Cells may also be implanted intracranially to overcome the blood–brain barrier[89]. Orthotopic homing and the metastatic potential of cancer cells can be boosted by genetic modification; for example, colorectal cancer PDX cells engineered to express C-C motif chemokine receptor 9 (CCR9) efficiently localize to the mouse colon after tail-vein injection, attracted by the abundance of the CCR9 ligand C-C motif chemokine ligand 25 (CCL25) in the intestine, and then spontaneously metastasize to the liver[90]. Genetic manipulation is useful to develop models of spontaneous metastasis for mechanistic studies *in vivo*; however, the introduction of exogenous molecules to patient-derived material may affect some properties of the original tumour, thus reducing translational relevance.

Whether PDOX models more accurately recapitulate clinical response to anticancer drugs compared with conventional subcutaneous PDX models remains to be established. One report showed that the antitumoural effects of a microtubule-stabilizing drug on PDX models of brain metastases from non-small-cell lung cancer were different in orthotopic versus subcutaneous implants[85], but results remain anecdotal. It is conceivable that therapies that target components of the tumour microenvironment, such as endothelial cells and immune cells, would be better evaluated in an orthotopic context. Conversely, the therapeutic response of ‘oncogene-addicted tumours’, which intrinsically rely on activating mutations for their growth and survival, is likely to be less dependent on anatomical location and more influenced by the underlying cancer genetic makeup. Indeed, despite their heterotopic location, subcutaneous PDXs from *BRAF*-mutant melanoma[9,91] and *HER2* (also known as *ERBB2*)-amplified colorectal cancer[6,92,93] mimic the therapeutic response observed in patients. Sharing results from different experimental models within the EurOPDX consortium will allow us to shed some light on this important question.

CTC-derived PDX models. As mentioned above, a step forwards for minimally invasive tumour sampling is the isolation and characterization of CTCs, detected at low concentrations in the peripheral blood of patients with different solid tumours[40]. Although the role of CTCs in metastasis development is still uncertain[40], their levels ostensibly correlate with patient survival and response to therapy[94–96]. These features mean that CTCs are promising tools to monitor cancer burden and drug susceptibility in metastatic and late-stage disease, when repetitive biopsies are not indicated. Technological advances now allow the isolation of viable CTCs, which maintain tumorigenicity when xenografted into immunocompromised mice[97–99] (TABLE 1).

Several reports have demonstrated the feasibility of establishing CTC-derived PDX models by directly injecting freshly isolated and enriched CTCs from patients with different cancers into immunocompromised mice. Using various CTC-capture techniques (such as epithelial cell adhesion molecule (EPCAM) or cytokeratin-based selection of cancer cells derived from epithelial tissues or microfluidic-based leukocyte depletion[100,101]), CTC-derived xenografts are now practicable for breast cancer[97], prostate cancer[102], gastric cancer[103], small-cell lung cancer (SCLC)[98] and melanoma[91]. Moreover, it has also been shown that *ex vivo* cultivated and fully molecularly characterized breast[104] and colorectal[105] CTCs maintain their tumorigenic potential. Notably, both freshly isolated CTCs and CTC-derived PDXs genetically and histologically mirror the original tumour and retain analogous drug sensitivities[91,97,98,100,102–105]. For example, PDXs that are established from chemotherapy-naive circulating SCLC cells recapitulate donor patients’ response to both platinum and etoposide[98]. In patients with ER-positive breast cancer, CTCs have also proved to be a useful model to study the genetic evolution of the tumour and to identify novel drug susceptibilities[104].

Although technically challenging, the use of CTC-derived PDX models opens new possibilities for translational research. In addition to being a source of information regarding disease prognosis[106], tumour heterogeneity[107,108], evolution[109] and dissemination[110,111], CTC-derived PDXs hold promise for precision medicine applications (TABLE 1). For example, CTCs from women with treatment-refractory ER-positive breast tumours have been recently analysed to investigate the functional and phenotypic consequences of prolonged anti-hormonal therapies, and xenografts from such CTCs

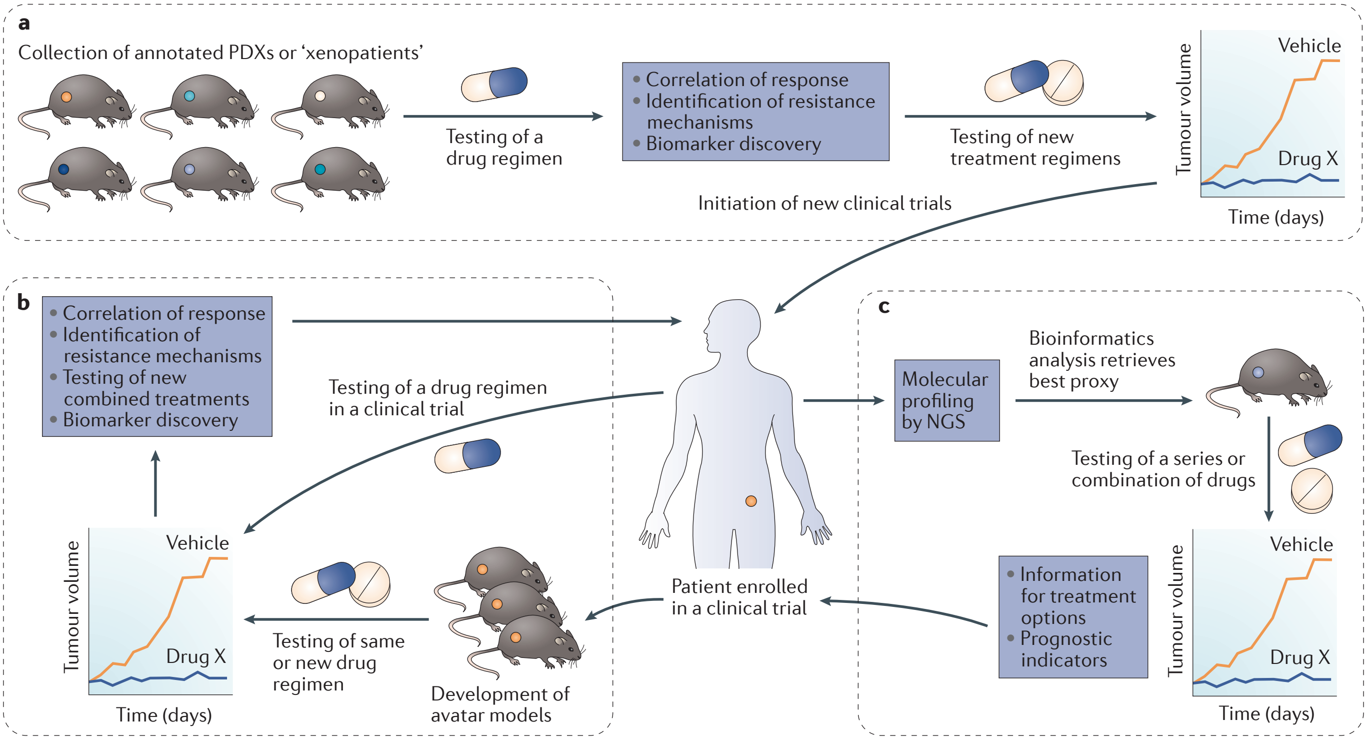

a
Collection of annotated PDXs or 'xenopatients'
Testing of a drug regimen
• Correlation of response
• Identification of resistance mechanisms
• Biomarker discovery
Testing of new treatment regimens
Tumour volume
Vehicle
Drug X
Time (days)
Initiation of new clinical trials
b
• Correlation of response
• Identification of resistance mechanisms
• Testing of new combined treatments
• Biomarker discovery
Testing of a drug regimen in a clinical trial
Tumour volume
Vehicle
Drug X
Time (days)
Testing of same or new drug regimen
Development of avatar models
Patient enrolled in a clinical trial
c
Molecular profiling by NGS
Bioinformatics analysis retrieves best proxy
Testing of a series or combination of drugs
• Information for treatment options
• Prognostic indicators
Tumour volume
Vehicle
Drug X
Time (days)

Figure 2 | **PDX preclinical study designs. a** | Large collections of patient-derived xenograft (PDX) models ('xenopatients') now allow population-based studies to be carried out, which better mimic the inter-tumour heterogeneity that is seen in patients and are more predictive of clinical efficacy than conventional xenografts of immortalized cancer cell lines. PDX molecular characterization and correlation with therapeutic response also facilitates biomarker discovery, as well as the identification of primary (and acquired) resistance mechanisms. These studies can lead to new hypotheses and support the initiation of new clinical trials. **b** | For some cancer types for which avatar models can be developed, co-clinical avatar studies allow for simultaneous drug testing in mice and patients for real-time adaptive therapeutic decisions. **c** | In the 'biofacsimile' or 'proxy' study format, integrative systems-based bioinformatics analysis can be used to pinpoint the best-matched PDX for a given patient from a collection of molecularly profiled models. PDX-associated information is then leveraged to instruct clinical treatment options and/or to derive prognostic indicators. NGS, next-generation sequencing.

have been used to design new therapies to overcome resistance[112]. Similarly, the next-generation sequencing of tumours, complemented with genomic analysis of CTCs and CTC-derived PDX mouse models, has proved to be a powerful platform for developing precision medicine strategies in patients with melanoma[91]. This approach has, in specific cases, facilitated the clinical implementation of alternative therapeutic strategies informed by the preclinical models[91].

PDXs for clinical decision-making

PDX population xenopatient trials. Across tumours of the same origin, genetic lesions that sustain tumorigenesis (and that therefore associate with response to targeted drugs) often involve many different oncogenes, each of which is mutated at a low frequency[113]. Furthermore, genotype-based prediction of drug response is not unequivocal. Despite harbouring the genetic lesion that is known to correlate with drug response, many tumours do not regress owing to the presence of signals that compensate for target inhibition[114]. Collectively, this information indicates that the genetic selection of tumours for the application of targeted therapies requires representative study populations and suitable pharmacogenomic platforms.

Provided that they are generated in high numbers and extensively characterized at the molecular level, PDXs can act as a powerful resource for large-scale genotype–response correlations and therapeutic studies in genetically defined tumour subsets. Several recent studies testify to this potential; in late-stage colorectal cancer, for example, a systematic assessment of response to antibodies targeting epidermal growth factor receptor (EGFR) using PDX models ('xenopatients') derived from hundreds of individual tumours was coupled to candidate-gene or whole-exome sequencing analyses. Through this effort, several genetic determinants of resistance to EGFR blockade were discovered, including amplifications or mutations in genes encoding druggable kinases[6,7,115,116]. Similarly, more dynamic features such as expression changes in pro-survival genes and the activation of compensatory feedback loops during treatment were identified as mechanisms of tumour adaptation to EGFR family[117,118] or MEK[119] inhibition in colorectal cancer. The flexibility of PDXs also enabled preclinical testing of drug combinations in models displaying some of these resistance traits, with a permutation capability that was clearly beyond the number of testable hypotheses in humans (FIG. 2).

An analogous population-based drug screen has recently been carried out in more than 1,000 PDX models representing a wide range of solid cancers (the 'PDX Encyclopaedia')[9]. Some genetic hypotheses and biomarkers of drug sensitivity, which emerged from cultured cancer cell lines, were successfully validated in this large panel of PDX models (FIG. 2). Notably, experiments in PDXs also enabled the identification of therapeutic candidates that *in vitro* model systems failed to capture[9]. In all these studies, responses obtained in mice were highly consistent with responses in patients. For example, the distribution of tumour regression, disease stabilization and progression in colorectal cancer

Table 3 | **Comparative quantitative data of response rates in PDXs versus human patients**

Tumour type	Clinical question	Comparative response rates	
		PDXs	Patients
CRC*	Response to EGFR antibody monotherapy in genetically unselected CRC PDXs[6] or unselected chemorefractory patients with CRC[178]	• PR: 5 of 47 (10.6%) • SD: 14 of 47 (29.8%) • PD: 28 of 47 (59.6%)	• PR: 12 of 111 (10.8%) • SD: 24 of 111 (21.6%) • PD: 59 of 111 (53.2%) • Not evaluated: 16 of 111 (14.4%)
CRC*	• PDXs[118]: response to EGFR antibody monotherapy in KRAS, NRAS and BRAF wild-type models • Patients[179]: response to EGFR antibody plus chemotherapy in chemorefractory patients with KRAS, NRAS and BRAF wild-type CRC	• PR: 31 of 125 (24.8%) • SD: 60 of 125 (48%) • PD: 34 of 125 (27.2%)	• PR: 15 of 56 (26.8%) • SD: 29 of 56 (51.8%) • PD: 12 of 56 (21.4%)
NSCLC	Co-clinical trial, PDX versus donor patient[66]: response to EGFR small-molecule inhibitors in four representative cases of six established PDXs	• 1 PR • 1 SD • 2 PD	• 1 PR • 1 SD • 2 PD
Breast cancer	Co-clinical trial, PDX versus donor patient[63]: response to several therapies	• Doxorubicin: 4 PD • Docetaxel: 1 PR and 6 PD • Anti-HER2 combination therapy (trastuzumab and lapatinib): 1 PR	• Doxorubicin: 4 PD • Docetaxel: 1 PR and 6 PD • Anti-HER2 combination therapy (trastuzumab and lapatinib): 1 PR

CRC, colorectal cancer; EGFR, epidermal growth factor receptor; NSCLC, non-small-cell lung cancer; PD, progressive disease; PDX, patient-derived xenograft; PR, partial response; SD, stable disease. *Data represent separate PDX and patient population studies.

xenopatients receiving EGFR antibodies was similar to that found in the clinic, and treatment-refractory tumour grafts were enriched for known genetic predictors of therapeutic resistance in patients[6] (TABLE 3); moreover, in analogy with clinical studies[120], the addition of an EGFR small-molecule inhibitor to the EGFR antibody increased tumour regression[118]. Similarly, PDXs from *BRAF*-mutant melanomas underwent substantial shrinkage when treated with BRAF inhibitors, a response that was further magnified – as in patients – by the addition of a MEK inhibitor[9,121]. PDX platforms have recently been used for the systematic identification of cancer vulnerabilities through RNA interference-based genetic screens in tumour grafts, which have revealed new oncogenic drivers in melanoma[122] and pancreatic tumours[123].

PDX population trials may be highly informative, but they are also expensive and technically cumbersome, and the trade-off between sufficient sample size to ensure adequate coverage of inter-patient heterogeneity and experimental feasibility requires careful study design. To reduce the number of animal replicates while preserving statistical power, reproducibility studies have been conducted to compare response calls made on a single mouse with majority responses in reference cohorts composed of many animals. Thus, a strong concordance between single-mouse responses and majority responses has been found, with a prediction accuracy varying from 75%[124] to 95%[9]. Accordingly, 'one animal per model per treatment' (1 × 1 × 1) approaches have recently been advocated[9,125].

Alternative strategies to reduce experimental burden could rely on step-wise approaches, in which large-scale pharmacogenomic screens are carried out using less laborious formats (such as cancer cell lines) followed by *in vivo* validation in selected, molecularly relevant PDX models. In this regard, it is noteworthy that patient-derived material from human tumours, such as colorectal, pancreas and prostate cancers[126–132], can be grown and nearly indefinitely expanded as three-dimensional (3D) organoids. These can be easily transplanted to establish PDXs, and vice versa, and are amenable to drug screens in a semi-high-throughput manner[130]. Albeit more difficult to establish and propagate, two-dimensional (2D) primary cultures of dissociated cancer cells from both patient samples and PDXs are also being attempted with a similar rationale and objectives[133]. In this vein, a platform for drug testing in short-term cultured breast cancer cells from PDXs has recently been developed and shown to predict *in vivo* drug response[29].

PDX co-clinical avatar trials. The term co-clinical trial refers to simultaneous clinical and preclinical trials with anticancer agents in patients with a tumour type of a defined genetic makeup and a mouse model with similar genetic abnormalities[134]. The underpinning idea is that the comparison of responses between the patients and the preclinical model will help to define the mechanism of action of a given drug, as well as biomarkers of response. Originally implemented with GEM models, the co-clinical trial concept has been expanded to include PDX models ('avatars'), which are generated from cancer patients enrolled in clinical trials and, in parallel, treated with the same drug or drugs that the patient is receiving[10] (FIG. 2). In general, these studies aim to develop a PDX model from newly diagnosed patients and use it to explore therapies that can be administered to the patient at the time of disease progression. Ongoing trials cover different tumour settings, including sarcomas (NCT02720796)[135], head and neck carcinomas (NCT02752932)[136], ovarian cancer (NCT02312245)[137] and pancreatic cancer (NCT02795650)[138]. Although a cogent argument exists for implementing avatar trials, and several case reports have provided data to support the concept[139–141], the logistical difficulties and technical hurdles are likely to limit the broad applicability of this approach (see above).

PDX models in biomarker development. The validation of mechanisms that link specific biomarkers to treatment efficacy will have direct clinical effects, allowing patient stratification for tailored treatment protocols. Large-scale PDX trial formats, such as the PDX Encyclopaedia[9] mentioned above, represent a more accurate approach to identify predictive biomarkers compared with the use of cell lines (TABLE 1).

A transcriptional profiling study on 85 PDX models of nine different cancer types treated with nine separate cancer drugs identified 1,578 genes, the expression of which correlated with sensitivity to at least one drug; 333 of these genes showed significant association with sensitivity to two or more drugs, and 32 genes predicted response to six or seven drugs[142]. This type of study provides an initial set of biomarkers that require further evaluation in clinical material to determine translatability into a clinically useful assay.

Epigenetic biomarkers, such as DNA methylation, can also be assessed in PDXs as possible response predictors. A study that included 28 glioblastoma PDOXs showed that the poly(ADP-ribose) polymerase (PARP) inhibitor veliparib significantly enhances the efficacy of temozolomide (TMZ) chemotherapy only in models with *O*-6-methylguanine-DNA methyltransferase (*MGMT*) promoter hyper-methylation[143]. On the basis of these data, *MGMT* promoter hyper-methylation was included as an eligibility criterion for TMZ and veliparib combination treatment in an ongoing phase II/III glioblastoma clinical trial (NCT02152982)[144].

Determinants of therapeutic sensitivity can be identified at the protein level using pathway analysis in PDXs: a proteomic survey of 20 PDX models of glioblastoma and their parental tumours identified a subset of cases with comparable proteomic profiles displaying high levels of expression and phosphorylation of EGFR and its downstream signalling proteins[145]. The expression and phosphorylation status of EGFR and downstream targets might be used as a predictive biomarker of response to EGFR inhibition in preclinical trials and, if successful, included in future clinical trials aiming to inhibit EGFR signalling in patients with glioblastoma.

PDX models are also useful for the preclinical identification of metabolic biomarkers using magnetic resonance spectroscopy (MRS). This technique has recently been used to demonstrate differences in metabolic characteristics between molecular subtypes of breast cancer[146,147]. Elevated phosphocholine levels and low glycerophosphocholine levels have been proposed to be metabolic markers of aggressive disease in breast cancer based on *in vitro* studies[148]. However, MRS on intact tissue from PDX models of poor-prognosis basal-like breast cancer displays an inverted metabolic profile, with high glycerophosphocholine concentration rather than high phosphocholine concentration[146,147]. These observations suggest that proper tumour architecture, as maintained in PDXs, influences choline metabolism. Accordingly, a strong correlation between PDX models and clinical material was observed in the expression of genes that are involved in key metabolic pathways[146]. MRS technology also holds potential for *in vivo* non-invasive detection of metabolic biomarkers through tailored techniques such as 31P MRS or hyperpolarized 13C MRS[149,150]. Recently, a proof-of-principle study demonstrated the ability of *in vivo* MRS to distinguish basal-like from luminal-like breast cancer PDXs non-invasively using 31P MRS imaging[151].

For some cancer types, the ability of tumours to successfully engraft in mice can be considered per se as a surrogate biomarker of risk for disease progression. For example, in mammary tumours, the ability to generate stable tumour grafts significantly predicted reduced survival[8,152], and gene expression signatures associated with successful PDX engraftment correlated with worse survival outcome when tested in prognostically annotated data sets of triple-negative breast cancer[153]. Similarly, tumour grafts of pancreatic ductal adenocarcinoma displayed higher expression of metastasis-associated genes compared with samples that failed engraftment, and patient donors of successfully engrafted tumours had shorter survival[154].

It is now well established that human tumour stromal cells are replaced by mouse counterparts following engraftment[155]. As a consequence of this substitution, species-specific RNA sequencing-based expression profiling of PDXs offers a unique opportunity to distinguish mouse stroma-derived transcripts from human cancer cell-derived transcripts without the need to physically separate the two components before RNA extraction. Such analyses led to the identification of stromal-associated transcriptional signatures in colorectal cancer that are associated with poor prognosis and treatment resistance[156]. The negative prognostic significance of tumour stromal transcriptional signatures and their value for therapeutic decision-making and patient follow-up have also been described in other reports[157,158].

Challenges and opportunities

Ideal animal models for preclinical experimentation in oncology should fulfil several criteria: reflecting the diversity of cancer patients at the epidemiological and molecular levels; retaining, to the highest possible extent, the functional, phenotypic and genotypic characteristics of human tumours; faithfully predicting response to therapies, and recapitulating mechanisms of innate and acquired resistance; and allowing for experimental flexibility.

Although PDXs fulfil several of these criteria and can be further improved to meet additional requirements, certain inherent limitations remain difficult to address. A major obstacle is the necessity of using immunocompromised mice to circumvent xenograft rejection. This requirement hampers the use of current PDX models to assess immunotherapeutics. Although emerging humanization procedures are now expected to overcome some of the most important concerns (see Supplementary information S1 (table)), issues still remain with the incorporation of particular immune cell types, immune responses and lymphoid structures into these humanized models and with the eradication of xenogeneic GvHD. It is expected that the development of novel immune-deficient mice will take advantage of emerging technologies based on engineered nuclease enzymes for genome editing (such as transcription activator-like effector nuclease (TALEN) and CRISPR–Cas9). These modifications will include the replacement or introduction of combinations of human-specific cytokine receptors and adhesion molecules, as well as more comprehensive sets of HLA class I and HLA class II molecules.

As mentioned above, serial passaging of tumours leads to the substitution of human stroma by murine components, and mouse-derived cytokines and growth factors in some cases do not crossreact with receptors that are expressed by human (cancer) cells[159–162]. This makes the contribution of the tumour microenvironment to drug response difficult to assess in PDXs. Moreover, the lack of a species-compatible tumour stroma complicates the identification of pharmacodynamic markers of target inactivation for drugs that intercept cancer-related microenvironmental processes, such as angiogenesis and inflammation. Although mouse humanization procedures seek to reconstitute the human immune system, the replacement of stromal elements such as endothelial cells and fibroblasts with their human counterparts is currently daunting, if not unfeasible.

PDX-based efforts for cancer precision medicine also require adequate logistics, from proper archival biobanking to continuous propagation of live biospecimens, intensive animal experimentation and systematic integration of therapeutic results with high-content molecular annotation. The perception of this complexity and the awareness that resource sustainability cannot be maintained by individual academic laboratories have fuelled initiatives for creating and implementing shared large-scale PDX platforms, including the European EurOPDX resource, the US National Cancer Institute (NCI) repository of patient-derived models, the Public Repository of Xenografts (PRoXe), the Children's Oncology Group (COG) cell culture and xenograft repository, and the Pediatric Preclinical Testing Consortium (PPTC) (BOX 1).

When dealing with such large multi-institutional platforms, standardized methodological procedures should be carried out to ensure reproducibility and to streamline readouts so that they are interpretable across different laboratories (BOX 1). Further, therapeutic outcomes should be univocally deciphered and stringently interpreted. Retardation of tumour growth during therapy typically results in tumours that are smaller than controls at end point, but larger than they were before starting treatment; although this may well suggest that the therapy is biologically active (because it affects cancer cell proliferation), it is not an indication that the therapy is clinically effective; indeed, this kind of response would be clinically defined as 'disease progression' or, at best, 'disease stabilization'. In the EurOPDX experience, even manifest effects of tumour growth inhibition — as observed, for example, after blockade of MEK in PDXs of *KRAS*-mutant colorectal cancer[125] — did not translate into clinical benefit when analogous therapies were applied to patients[163]. By contrast, overt regression in PDXs predicted positive results in the clinic: the finding that an antibody and small molecule combination against HER2 induced massive regressions in *HER2*-amplified colorectal tumour grafts[6,117] has recently been translated into a successful clinical trial, with the vast majority of patients achieving tumour shrinkage when treated with the same regimen[93]. It has also become increasingly clear that the use of quantitative metrics to classify response (equivalent to clinical Response Evaluation Criteria in Solid Tumours (RECIST)) should be implemented to more precisely assess therapeutic effects in PDX trials. Modified RECIST criteria for mouse xenograft applications have recently been described[9]. 'Best response' is defined as the minimum value of percentage tumour volume change, compared with tumour volume at baseline, for treatment durations equal to or longer than 10 days, and 'best average response' is the minimum value of the mean percentage of tumour volume change, as measured at each evaluation time point along treatment, compared with baseline[9]. Such definitions, coupled with specific tumour volume cut-offs, have been applied to categorize complete response, partial response, stable disease and progressive disease in tumour-bearing mice. These modified RECIST criteria capture response kinetics, robustness and durability, and thus improve the ability of preclinical studies to accurately predict patient outcome.

Extended and detailed molecular annotation is a prerequisite for precision oncology paradigms. However, the accumulation of multiple layers of genomic information requires the development of computational systems with common or interoperable standards for normalization, correction and retrieval of complex data sets. The issue of big data collection, harmonization and storage is particularly important when working with large PDX collections, in which one original tumour from a single patient gives rise, upon serial passages, to many descendants that expand at an exponential rate (BOX 2). In EurOPDX, efforts are ongoing to aggregate cancer genomic profiles obtained through different technologies in different laboratories and to implement a user-friendly, open-source portal that showcases the molecular characteristics of the participating collections (BOX 1). Importantly, besides the detection of individual variants with clinically actionable potential, multi-dimensional molecular information from existing PDX models can be subjected to systems-based bioinformatics analysis to extract algorithms that identify key biological parameters[164]. Preliminary evidence suggests that such algorithms can

Box 2 | Data management and integration

By combining the flexibility of preclinical analysis with the instructive value of population-based studies, patient-derived xenografts (PDXs) offer unprecedented opportunities for drawing statistically robust correlations between genetic or functional traits and sensitivity to anticancer drugs. However, the advantages of high-throughput studies with PDX-based approaches may become major hurdles when dealing with large-scale data management, analysis and utilization. The deployment of PDX models for translational studies often requires their stratification into existing predictive or prognostic molecular classes and subgroups as derived on tumours from patients. The portability of the stratification criteria from human to mice, and vice versa, is not trivial, owing to multiple sources of biological and genomic variation, which may be introduced in the process of engrafting and propagating patient tumour material into murine hosts.

Data management issues

Data complexity and dynamics. The representation of cancer data in classical oncogenomic portals is normally static: the results obtained by analysing such public resources are not fed back to refine, update or complement the original information. The possibility to incrementally stratify and integrate multiple layers of information generated from the same original sample by diverse laboratories at different times represents one of the key added values of PDX-based approaches. This implies the need for further dimensions of complexity to interrogate an almost infinite number of variables and to implement decision-making algorithms in case of data inconsistency across experiments[166].

Data normalization and annotations. The joint utilization of human and PDX data requires the standardization of sample metadata such as clinical and molecular ontologies. Through this effort, data derived from different experiments, technologies and platforms can be normalized against common categories and used to interrogate samples with integrative queries exploring heterogeneous data domains.

Data analysis issues

Population selection bias. Owing to the different engraftment efficacies inherent to each tumour sample, the population of xenografts might not recapitulate the full distribution of tumour phenotypic or molecular variables observed in patients. Any prior-dependent statistical models should be adapted to the new distribution of subclasses within the PDX population. This implies the necessity to identify the missing or underrepresented subgroups through analytical investigation of multidimensional parameters (genomics, transcriptomics, histopathology, and so on).

Loss of human immune and stromal cells. Although both stromal and immune components are replaced over time by murine analogues, the haematopoietic elements show important differences in their spatial distribution[167] or may be missing overall[156,168]. This affects the signal received from molecular profiling, and could require the application of specific algorithms for signal correction to avoid or reduce artefacts and biases[156,169].

be subsequently used to identify one or more 'biofacsimile' or 'proxy' PDX models for individual patients, and PDX-associated information may be leveraged to instruct treatment options and/or to derive predictive indicators in the clinic[164] (FIG. 2).

All these considerations underscore the opportunities offered by PDX models to illuminate new angles of translational cancer research, but they also put forward the challenges that are intrinsic to this approach, and the need for finding new ways to maximize PDX potential. Industry-led PDX ventures rely on common and extensively tested operating procedures, backed by considerable funding, which ensures scalable, homogeneous and reproducible experimental schemes; however, pharmaceutical initiatives are typically bound to preclinical testing of proprietary compounds and may face obstacles in publishing results, especially when data relate to sensitive commercial or patenting issues. Conversely, owing to their multi-institutional nature, scholarly consortia usually suffer from heterogeneous characterization of their PDX collections, a flaw that is hardly corrected by the relatively limited resources provided by government or charity grants; however, PDX academic efforts enjoy flexibility in drug testing and unfettered scientific reporting (including reporting of negative results, which avoids the duplication of effort and reduces costs). As EurOPDX members working in academia, we share with our colleagues of PRoXe the concern that "academic centers are ill suited to bear the burden of housing, expanding, archiving, characterizing, and disseminating PDXs to investigators (academic and industrial) across the world" (REF. 165). Meanwhile, we believe that joining forces, incorporating models, coordinating methodologies, and improving the public shareability and visibility of molecular data in an academic-oriented rather than in an industry-scale format are viable objectives that will foster not only a stronger collaborative spirit in cancer medicine, but also a change of mind-set within institutional authorities and industrial stakeholders. EurOPDX started as a crowd-funded initiative of scientists with common goals, complementary skills and similar needs, and is now growing in a more structured manner thanks to enterprise-wide development plans. Ultimately, we envision a virtuous circle in which new knowledge from bottom-up efforts such as ours and others will inform clinical decision making, which in turn will orient public and private financial interests to secure further sustainability of PDX-based activities. Successful examples in other contexts of biomedical research, such as TRANSAUTOPHAGY (see Further information; a European consortium for multidisciplinary research and translation of knowledge on autophagy) and GENiE (see Further information; a network of scientists using *Caenorhabditis elegans* as a model organism), bode well to achieve this ambition.

Annette T. Byrne and Monika A. Jarzabek are members of the EurOPDX Consortium and are at the Royal College of Surgeons in Ireland, Dublin 2, Ireland.

Denis G. Alférez and Robert B. Clarke are members of the EurOPDX Consortium and are at the Breast Cancer Now Research Unit, Division of Molecular and Clinical Cancer Sciences, Manchester Cancer Research Centre, University of Manchester, Manchester M20 4QL, UK.

Frédéric Amant, Daniela Annibali and Els Hermans are members of the EurOPDX Consortium and are at the Katholieke Universiteit Leuven, 3000 Leuven, Belgium. Frédéric Amant is also at The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands.

Joaquín Arribas, Joan Seoane and Laura Soucek are members of the EurOPDX Consortium and are at the Vall d'Hebron Institute of Oncology, 08035 Barcelona, the Universitat Autònoma de Barcelona, 08193 Bellaterra, and the Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. Joaquín Arribas and Joan Seoane are also at CIBERONC, 08035 Barcelona, Spain.

Andrew V. Biankin and David K. Chang are members of the EurOPDX Consortium and are at the Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Glasgow G61 1QH, UK.

Alejandra Bruna, Carlos Caldas and Oscar M. Rueda are members of the EurOPDX Consortium and are at Cancer Research UK Cambridge Institute, Cambridge Cancer Centre, University of Cambridge, Cambridge CB2 0RE, UK.

Eva Budinská is a member of the EurOPDX Consortium and is at the Institute of Biostatistics and Analyses, Faculty of Medicine, and Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masarykova Univerzita, 625 00 Brno, Czech Republic.

Hans Clevers is at the Hubrecht Institute, University Medical Centre Utrecht, and Princess Maxima Center for Pediatric Oncology, 3584CT Utrecht, The Netherlands.

George Coukos and Dominique Vanhecke are members of the EurOPDX Consortium and are at Lausanne Branch, Ludwig Institute for Cancer Research at the University of Lausanne, 1066 Lausanne, Switzerland.

Virginie Dangles-Marie is a member of the EurOPDX Consortium and is at the Institut Curie, PSL Research University, Translational Research Department, 75005 Paris, and Université Paris Descartes, Sorbonne Paris Cité, Faculté de Pharmacie de Paris, 75006 Paris, France.

S. Gail Eckhardt is at the University of Colorado Cancer Center, Aurora, Colorado 80045, USA.

Eva Gonzalez-Suarez is a member of the EurOPDX Consortium and is at the Cancer Epigenetics and Biology Program, Bellvitge Biomedical Research Institute IDIBELL, 08908 L'Hospitalet de Llobregat, Barcelona, Spain.

Manuel Hidalgo is a member of the EurOPDX Consortium and is at Beth Israel Deaconess Medical Center, Boston, Harvard Medical School, Boston, Massachusetts 02215, USA.

Steven de Jong is a member of the EurOPDX Consortium and is at the University Medical Centre Groningen, University of Groningen, 9713GZ Groningen, The Netherlands.

Jos Jonkers, Kristel Kemper and Daniel S. Peeper are members of the EurOPDX Consortium and are at The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands.

Luisa Lanfrancone and Pier Giuseppe Pelicci are members of the EurOPDX Consortium and are at the Department of Experimental Oncology, European Instiitute of Oncology, 20139 Milan, Italy.

Gunhild Mari Mælandsmo and Jens Henrik Norum are members of the EurOPDX Consortium and are at Oslo University Hospital, Institute for Cancer Research, 0424 Oslo, Norway.

Elisabetta Marangoni and Sergio Roman-Roman are members of the EurOPDX Consortium and are at Institut Curie, PSL Research University, Translational Research Department, 75005 Paris, France.

Jean-Christophe Marine is a member of the EurOPDX Consortium and is at the Laboratory for Molecular Cancer Biology, Department of Oncology, Katholieke Universiteit Leuven, and the Center for Cancer Biology, VIB, 3000 Leuven, Belgium.

Enzo Medico, Andrea Bertotti and Livio Trusolino are members of the EurOPDX Consortium and are at the Candiolo Cancer Institute IRCCS and Department of Oncology, University of Torino, 10060 Candiolo, Torino, Italy.

Héctor G. Palmer, Alejandro Piris-Gimenez and Violeta Serra are members of the EurOPDX Consortium and are at the Vall d'Hebron Institute of Oncology and CIBERONC, 08035 Barcelona, Spain.

Alberto Villanueva is a member of the EurOPDX Consortium and is at the Program Against Cancer Therapeutic Resistance (ProCURE), Catalan Institute of Oncology ICO, Bellvitge Biomedical Research Institute IDIBELL, 08098 L'Hospitalet de Llobregat, Barcelona, and Xenopat S.L., Business Bioincubator, Bellvitge Health Science Campus, 08907 L'Hospitalet de Llobregat, Barcelona, Spain.

Emilie Vinolo is at Seeding Science SAS, 75020 Paris, France.

Correspondence to A.T.B. and L.T.
annettebyrne@rcsi.ie;
livio.trusolino@ircc.it

doi:10.1038/nrc.2016.140
Published online 20 Jan 2017

1. de Bono, J. S. & Ashworth, A. Translating cancer research into targeted therapeutics. *Nature* **467**, 543–549 (2010).
2. Daniel, V. C. *et al.* A primary xenograft model of small-cell lung cancer reveals irreversible changes in gene expression imposed by culture *in vitro*. *Cancer Res.* **69**, 3364–3373 (2009).
3. Arrowsmith, J. Trial watch: phase II failures: 2008–2010. *Nat. Rev. Drug Discov.* **10**, 328–329 (2011).
4. Arrowsmith, J. & Miller, P. Trial watch: phase II and phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* **12**, 569 (2013).
5. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
6. Bertotti, A. *et al.* A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-

resistant colorectal cancer. *Cancer Discov.* **1**, 508–523 (2011).

7. Bertotti, A. *et al.* The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature* **526**, 263–267 (2015).
8. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17**, 1514–1520 (2011).
9. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
10. Hidalgo, M. *et al.* Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* **4**, 998–1013 (2014).
11. Siolas, D. & Hannon, G. J. Patient-derived tumor xenografts: transforming clinical samples into mouse models. *Cancer Res.* **73**, 5315–5319 (2013).
12. Tentler, J. J. *et al.* Patient-derived tumour xenografts as models for oncology drug development. *Nat. Rev. Clin. Oncol.* **9**, 338–350 (2012).
13. Day, C. P., Merlino, G. & Van Dyke, T. Preclinical mouse cancer models: a maze of opportunities and challenges. *Cell* **163**, 39–53 (2015).
14. Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. *Nat. Rev. Cancer* **15**, 473–483 (2015).
15. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368**, 842–851 (2013).
16. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep.* **6**, 514–527 (2014).
17. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
18. Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543–548 (2013).
19. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer* **12**, 323–334 (2012).
20. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).
21. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
22. Dawson, S. J., Rueda, O. M., Aparicio, S. & Caldas, C. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J.* **32**, 617–628 (2013).
23. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
24. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
25. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
26. Bhang, H. E. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* **21**, 440–448 (2015).
27. Jeselsohn, R. *et al.* Emergence of constitutively active estrogen receptor-α mutations in pretreated advanced estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **20**, 1757–1767 (2014).
28. Murtaza, M. *et al.* Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.* **6**, 8760 (2015).
29. Bruna, A. *et al.* A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell* **167**, 1–15 (2016).
30. Marangoni, E. *et al.* A new model of patient tumor-derived breast cancer xenografts for preclinical assays. *Clin. Cancer Res.* **13**, 3989–3998 (2007).
31. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4**, 1116–1130 (2013).
32. Cassidy, J. W., Caldas, C. & Bruna, A. Maintaining tumor heterogeneity in patient-derived tumor xenografts. *Cancer Res.* **75**, 2963–2968 (2015).
33. Cottu, P. *et al.* Acquired resistance to endocrine treatments is associated with tumor-specific molecular changes in patient-derived luminal breast cancer xenografts. *Clin. Cancer Res.* **20**, 4314–4325 (2014).
34. Ter Brugge, P. *et al.* Mechanisms of therapy resistance in patient-derived xenograft models of BRCA1-deficient breast cancer. *J. Natl Cancer Inst.* **108**, djw148 (2016).
35. Kemper, K. *et al.* Intra- and inter-tumor heterogeneity in a vemurafenib-resistant melanoma patient and derived xenografts. *EMBO Mol. Med.* **7**, 1104–1118 (2015).
36. Shi, H. *et al.* Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discov.* **4**, 80–93 (2014).
37. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
38. Kemper, K. *et al.* $BRAF^{V600E}$ kinase domain duplication identified in therapy-refractory melanoma patient-derived xenografts. *Cell Rep.* **16**, 263–277 (2016).
39. Nguyen, L. V. *et al.* DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nat. Commun.* **5**, 5871 (2014).
40. Joosse, S. A., Gorges, T. M. & Pantel, K. Biology, detection, and clinical implications of circulating tumor cells. *EMBO Mol. Med.* **7**, 1–11 (2015).
41. Massague, J. & Obenauf, A. C. Metastatic colonization by circulating tumour cells. *Nature* **529**, 298–306 (2016).
42. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
43. Pece, S. *et al.* Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell* **140**, 62–73 (2010).
44. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
45. Li, C., Lee, C. J. & Simeone, D. M. Identification of human pancreatic cancer stem cells. *Methods Mol. Biol.* **568**, 161–173 (2009).
46. Lawson, D. A. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).
47. Li, X. *et al.* Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J. Natl Cancer Inst.* **100**, 672–679 (2008).
48. Todaro, M. *et al.* Colon cancer stem cells dictate tumor growth and resist cell death by production of interleukin-4. *Cell Stem Cell* **1**, 389–402 (2007).
49. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
50. Fan, F. *et al.* The requirement for freshly isolated human colorectal cancer (CRC) cells in isolating CRC stem cells. *Br. J. Cancer* **112**, 539–546 (2015).
51. Borovski, T., De Sousa, E. M. F., Vermeulen, L. & Medema, J. P. Cancer stem cell niche: the place to be. *Cancer Res.* **71**, 634–639 (2011).
52. Charafe-Jauffret, E. *et al.* ALDH1-positive cancer stem cells predict engraftment of primary breast tumors and are governed by a common stem cell program. *Cancer Res.* **73**, 7290–7300 (2013).
53. Miranda-Lorenzo, I. *et al.* Intracellular autofluorescence: a biomarker for epithelial cancer stem cells. *Nat. Methods* **11**, 1161–1169 (2014).
54. Sainz, B. Jr *et al.* Microenvironmental hCAP-18/LL-37 promotes pancreatic ductal adenocarcinoma by activating its cancer stem cell compartment. *Gut* **64**, 1921–1935 (2015).
55. Magee, J. A., Piskounova, E. & Morrison, S. J. Cancer stem cells: impact, heterogeneity, and uncertainty. *Cancer Cell* **21**, 283–296 (2012).
56. Rottenberg, S. *et al.* Selective induction of chemotherapy resistance of mammary tumors in a conditional mouse model for hereditary breast cancer. *Proc. Natl Acad. Sci. USA* **104**, 12117–12122 (2007).
57. Castillo-Avila, W. *et al.* Sunitinib inhibits tumor growth and synergizes with cisplatin in orthotopic models of cisplatin-sensitive and cisplatin-resistant human testicular germ cell tumors. *Clin. Cancer Res.* **15**, 3384–3395 (2009).
58. Juliachs, M. *et al.* The PDGFRβ–AKT pathway contributes to CDDP-acquired resistance in testicular germ cell tumors. *Clin. Cancer Res.* **20**, 658–667 (2014).
59. Simoes, B. M. *et al.* Anti-estrogen resistance in human breast tumors is driven by JAG1-NOTCH4-dependent cancer stem cell activity. *Cell Rep.* **12**, 1968–1977 (2015).
60. Herrera-Abreu, M. T. *et al.* Early adaptation and acquired resistance to CDK4/6 inhibition in estrogen receptor-positive breast cancer. *Cancer Res.* **76**, 2301–2313 (2016).
61. Kim, K. T. *et al.* Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* **16**, 127 (2015).
62. Cottu, P. *et al.* Modeling of response to endocrine therapy in a panel of human luminal breast cancer xenografts. *Breast Cancer Res. Treat.* **133**, 595–606 (2012).
63. Zhang, X. *et al.* A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73**, 4885–4897 (2013).
64. Das Thakur, M. *et al.* Modelling vemurafenib resistance in melanoma reveals a strategy to forestall drug resistance. *Nature* **494**, 251–255 (2013).
65. Sun, C. *et al.* Reversible and adaptive resistance to $BRAF^{V600E}$ inhibition in melanoma. *Nature* **508**, 118–122 (2014).
66. Stewart, E. L. *et al.* Clinical utility of patient-derived xenografts to determine biomarkers of prognosis and map resistance pathways in EGFR-mutant lung adenocarcinoma. *J. Clin. Oncol.* **33**, 2472–2480 (2015).
67. Stebbing, J. *et al.* Patient-derived xenografts for individualized care in advanced sarcoma. *Cancer* **120**, 2006–2015 (2014).
68. Balko, J. M. *et al.* Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. *Cancer Discov.* **4**, 232–245 (2014).
69. Zacarias-Fluck, M. F. *et al.* Effect of cellular senescence on the growth of HER2-positive breast cancers. *J. Natl Cancer Inst.* **107**, djv020 (2015).
70. Bankert, R. B., Egilmez, N. K. & Hess, S. D. Human-SCID mouse chimeric models for the evaluation of anti-cancer therapies. *Trends Immunol.* **22**, 386–393 (2001).
71. Hylander, B. L. *et al.* Origin of the vasculature supporting growth of primary patient tumor xenografts. *J. Transl Med.* **11**, 110 (2013).
72. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).
73. Guichelaar, T. *et al.* Human regulatory T cells do not suppress the antitumor immunity in the bone marrow: a role for bone marrow stromal cells in neutralizing regulatory T cells. *Clin. Cancer Res.* **19**, 1467–1475 (2013).
74. King, M. A. *et al.* Human peripheral blood leucocyte non-obese diabetic-severe combined immunodeficiency interleukin-2 receptor gamma chain gene mouse model of xenogeneic graft-versus-host-like disease and the role of host major histocompatibility complex. *Clin. Exp. Immunol.* **157**, 104–118 (2009).
75. Holzapfel, B. M., Wagner, F., Thibaudeau, L., Levesque, J. P. & Hutmacher, D. W. Concise review: humanized models of tumor immunology in the 21st century: convergence of cancer research and tissue engineering. *Stem Cells* **33**, 1696–1704 (2015).
76. Drake, A. C., Chen, Q. & Chen, J. Engineering humanized mice for improved hematopoietic reconstitution. *Cell. Mol. Immunol.* **9**, 215–224 (2012).
77. Reinisch, A., Gratzinger, D., Hong, W.-J. & Majeti, R. A. Novel humanized bone marrow niche xenotransplantation model allows superior engraftment of human normal and malignant hematopoietic cells and reveals myelofibrosis-initiating cells in the HSC compartment. *Blood* **124**, 349 (2014).
78. Rongvaux, A. *et al.* Human hemato-lymphoid system mice: current use and future potential for medicine. *Annu. Rev. Immunol.* **31**, 635–674 (2013).
79. Voloshin, T. *et al.* G-CSF supplementation with chemotherapy can promote revascularization and subsequent tumor regrowth: prevention by a CXCR4 antagonist. *Blood* **118**, 3426–3435 (2011).
80. Morton, J. J. *et al.* XactMice: humanizing mouse bone marrow enables microenvironment reconstitution in a patient-derived xenograft model of head and neck cancer. *Oncogene* **35**, 290–300 (2016).
81. Takenaka, K. *et al.* Polymorphism in Sirpa modulates engraftment of human hematopoietic stem cells. *Nat. Immunol.* **8**, 1313–1323 (2007).
82. Du, Q. *et al.* Establishment of and comparison between orthotopic xenograft and subcutaneous xenograft models of gallbladder carcinoma. *Asian Pac. J. Cancer Prev.* **15**, 3747–3752 (2014).

83. Hoffman, R. M. Patient-derived orthotopic xenografts: better mimic of metastasis than subcutaneous xenografts. *Nat. Rev. Cancer* **15**, 451–452 (2015).
84. Dai, L., Lu, C., Yu, X. I., Dai, L. J. & Zhou, J. X. Construction of orthotopic xenograft mouse models for human pancreatic cancer. *Exp. Ther. Med.* **10**, 1033–1038 (2015).
85. Ambrogio, C. *et al.* Combined inhibition of DDR1 and Notch signaling is a therapeutic strategy for KRAS-driven lung adenocarcinoma. *Nat. Med.* **22**, 270–277 (2016).
86. de Jong, M., Essers, J. & van Weerden, W. M. Imaging preclinical tumour models: improving translational power. *Nat. Rev. Cancer* **14**, 481–493 (2014).
87. Iorns, E. *et al.* A new mouse model for the study of human breast cancer metastasis. *PLoS ONE* **7**, e47995 (2012).
88. Gupta, P., Adkins, C., Lockman, P. & Srivastava, S. K. Metastasis of breast tumor cells to brain is suppressed by phenethyl isothiocyanate in a novel metastasis model. *PLoS ONE* **8**, e67278 (2013).
89. Lee, H. W. *et al.* Patient-derived xenografts from non-small cell lung cancer brain metastases are valuable translational platforms for the development of personalized targeted therapy. *Clin. Cancer Res.* **21**, 1172–1182 (2015).
90. Chen, H. J. *et al.* Comprehensive models of human primary and metastatic colorectal tumors in immunodeficient and immunocompetent mice by chemokine targeting. *Nat. Biotechnol.* **33**, 656–660 (2015).
91. Girotti, M. R. *et al.* Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma. *Cancer Discov.* **6**, 286–299 (2016).
92. Nunes, M. *et al.* Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data. *Cancer Res.* **75**, 1560–1566 (2015).
93. Sartore-Bianchi, A. *et al.* Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. *Lancet Oncol.* **17**, 738–746 (2016).
94. Krebs, M. G. *et al.* Evaluation and prognostic significance of circulating tumor cells in patients with non-small-cell lung cancer. *J. Clin. Oncol.* **29**, 1556–1563 (2011).
95. Scher, H. I. *et al.* Circulating tumour cells as prognostic markers in progressive, castration-resistant prostate cancer: a reanalysis of IMMC38 trial data. *Lancet Oncol.* **10**, 233–239 (2009).
96. Zhang, L. *et al.* Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. *Clin. Cancer Res.* **18**, 5701–5710 (2012).
97. Baccelli, I. *et al.* Identification of a population of blood circulating tumor cells from breast cancer patients that initiates metastasis in a xenograft assay. *Nat. Biotechnol.* **31**, 539–544 (2013).
98. Hodgkinson, C. L. *et al.* Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. *Nat. Med.* **20**, 897–903 (2014).
99. Yap, T. A., Lorente, D., Omlin, A., Olmos, D. & de Bono, J. S. Circulating tumor cells: a multifunctional biomarker. *Clin. Cancer Res.* **20**, 2553–2568 (2014).
100. Alix-Panabieres, C. & Pantel, K. Challenges in circulating tumour cell research. *Nat. Rev. Cancer* **14**, 623–631 (2014).
101. Ignatiadis, M., Lee, M. & Jeffrey, S. S. Circulating tumor cells and circulating tumor DNA: challenges and opportunities on the path to clinical utility. *Clin. Cancer Res.* **21**, 4786–4800 (2015).
102. Williams, E. S. *et al.* Generation of prostate cancer patient derived xenograft models from circulating tumor cells. *J. Vis. Exp.* **104**, e53182 (2015).
103. Toyoshima, K. *et al.* Analysis of circulating tumor cells derived from advanced gastric cancer. *Int. J. Cancer* **137**, 991–998 (2015).
104. Yu, M. *et al.* Cancer therapy. *Ex vivo* culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science* **345**, 216–220 (2014).
105. Cayrefourcq, L. *et al.* Establishment and characterization of a cell line from human circulating colon cancer cells. *Cancer Res.* **75**, 892–901 (2015).
106. Aggarwal, C. *et al.* Relationship among circulating tumor cells, CEA and overall survival in patients with metastatic colorectal cancer. *Ann. Oncol.* **24**, 420–428 (2013).
107. Vishnoi, M. *et al.* The isolation and characterization of CTC subsets related to breast cancer dormancy. *Sci. Rep.* **5**, 17533 (2015).
108. Krebs, M. G. *et al.* Molecular analysis of circulating tumour cells-biology and biomarkers. *Nat. Rev. Clin. Oncol.* **11**, 129–144 (2014).
109. Markou, A. *et al.* PIK3CA mutational status in circulating tumor cells can change during disease recurrence or progression in patients with breast cancer. *Clin. Cancer Res.* **20**, 5823–5834 (2014).
110. Giuliano, M. *et al.* Circulating and disseminated tumor cells from breast cancer patient-derived xenograft-bearing mice as a novel model to study metastasis. *Breast Cancer Res.* **17**, 3 (2015).
111. Torphy, R. J. *et al.* Circulating tumor cells as a biomarker of response to treatment in patient-derived xenograft mouse models of pancreatic adenocarcinoma. *PLoS ONE* **9**, e89474 (2014).
112. Jordan, N. V. *et al.* HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature* **537**, 102–106 (2016).
113. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
114. Trusolino, L. & Bertotti, A. Compensatory pathways in oncogenic kinase signaling and resistance to targeted therapies: six degrees of separation. *Cancer Discov.* **2**, 876–880 (2012).
115. Bardelli, A. *et al.* Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer. *Cancer Discov.* **3**, 658–673 (2013).
116. Kavuri, S. M. *et al.* HER2 activating mutations are targets for colorectal cancer treatment. *Cancer Discov.* **5**, 832–841 (2015).
117. Leto, S. M. *et al.* Sustained inhibition of HER3 and EGFR is necessary to induce regression of HER2-amplified gastrointestinal carcinomas. *Clin. Cancer Res.* **21**, 5519–5531 (2015).
118. Zanella, E. R. *et al.* IGF2 is an actionable target that identifies a distinct subpopulation of colorectal cancer patients with marginal response to anti-EGFR therapies. *Sci. Transl Med.* **7**, 272ra12 (2015).
119. Sun, C. *et al.* Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. *Cell Rep.* **7**, 86–93 (2014).
120. Weickhardt, A. J. *et al.* Dual targeting of the epidermal growth factor receptor using the combination of cetuximab and erlotinib: preclinical evaluation and results of the phase II DUX study in chemotherapy-refractory, advanced colorectal cancer. *J. Clin. Oncol.* **30**, 1505–1512 (2012).
121. Long, G. V. *et al.* Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *N. Engl. J. Med.* **371**, 1877–1888 (2014).
122. Bossi, D. *et al. In vivo* genetic screens of patient-derived tumors revealed unexpected frailty of the transformed phenotype. *Cancer Discov.* **6**, 650–663 (2016).
123. Carugo, A. *et al. In vivo* functional platform targeting patient-derived xenografts identifies WDR5-Myc association as a critical determinant of pancreatic cancer. *Cell Rep.* **16**, 133–147 (2016).
124. Murphy, B. *et al.* Evaluation of alternative *in vivo* drug screening methodology: a single mouse analysis. *Cancer Res.* **76**, 5798–5809 (2016).
125. Migliardi, G. *et al.* Inhibition of MEK and PI3K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clin. Cancer Res.* **18**, 2515–2525 (2012).
126. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. *Cell* **160**, 324–338 (2015).
127. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. *Cell* **159**, 176–187 (2014).
128. Huang, L. *et al.* Ductal pancreatic cancer modeling and drug screening using human pluripotent stem cell-and patient-derived tumor organoids. *Nat. Med.* **21**, 1364–1371 (2015).
129. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
130. van de Wetering, M. *et al.* Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161**, 933–945 (2015).
131. Weeber, F. *et al.* Preserved genetic diversity in organoids cultured from biopsies of human colorectal cancer metastases. *Proc. Natl Acad. Sci. USA* **112**, 13308–13311 (2015).
132. Hubert, C. G. *et al.* A three-dimensional organoid culture system derived from human glioblastomas recapitulates the hypoxic gradients and cancer stem cell heterogeneity of tumors found *in vivo*. *Cancer Res.* **76**, 2465–2477 (2016).
133. Crystal, A. S. *et al.* Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science* **346**, 1480–1486 (2014).
134. Nardella, C., Lunardi, A., Patnaik, A., Cantley, L. C. & Pandolfi, P. P. The APL paradigm and the "co-clinical trial" project. *Cancer Discov.* **1**, 108–116 (2011).
135. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02720796 (2016).
136. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02752932 (2016).
137. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02312245 (2016).
138. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02795650 (2016).
139. Azaro, A. *et al.* A first-in-human phase I trial of LY2780301, a dual p70 S6 kinase and Akt Inhibitor, in patients with advanced or metastatic cancer. *Invest. New Drugs* **33**, 710–719 (2015).
140. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. *Nature* **518**, 240–244 (2015).
141. Morelli, M. P. *et al.* Prioritizing phase I treatment options through preclinical testing on personalized tumorgraft. *J. Clin. Oncol.* **30**, e45–e48 (2012).
142. Zembutsu, H. *et al.* Genome-wide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. *Cancer Res.* **62**, 518–527 (2002).
143. Gupta, S. K. *et al.* Delineation of MGMT Hypermethylation as a biomarker for veliparib-mediated temozolomide-sensitizing therapy of glioblastoma. *J. Natl Cancer Inst.* **108**, djv369 (2016).
144. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02152982 (2016).
145. Brown, K. E. *et al.* Proteomic profiling of patient-derived glioblastoma xenografts identifies a subset with activated EGFR: implications for drug development. *J. Neurochem.* **133**, 730–738 (2015).
146. Grinde, M. T. *et al.* Interplay of choline metabolites and genes in patient-derived breast cancer xenografts. *Breast Cancer Res.* **16**, R5 (2014).
147. Moestue, S. A. *et al.* Distinct choline metabolic profiles are associated with differences in gene expression for basal-like and luminal-like breast cancer xenograft models. *BMC Cancer* **10**, 433 (2010).
148. Glunde, K., Jie, C. & Bhujwalla, Z. M. Molecular causes of the aberrant choline phospholipid metabolism in breast cancer. *Cancer Res.* **64**, 4270–4276 (2004).
149. Nelson, S. J. *et al.* Metabolic imaging of patients with prostate cancer using hyperpolarized [1-^{13}C]pyruvate. *Sci. Transl. Med.* **5**, 198ra108 (2013).
150. Klomp, D. W. *et al.* 31P MRSI and 1H MRS at 7 T: initial results in human breast cancer. *NMR Biomed.* **24**, 1337–1342 (2011).
151. Esmaeili, M. *et al. In vivo* ^{31}P magnetic resonance spectroscopic imaging (MRSI) for metabolic profiling of human breast cancer xenografts. *J. Magn. Reson. Imaging* **41**, 601–609 (2015).
152. Eyre, R. *et al.* Patient-derived mammosphere and xenograft tumour initiation correlates with progression to metastasis. *J. Mammary Gland Biol. Neoplasia* http://dx.doi.org/10.1007/s10911-016-9361-8 (2016).
153. Moon, H. G. *et al.* Prognostic and functional importance of the engraftment-associated genes in the patient-derived xenograft models of triple-negative breast cancers. *Breast Cancer Res. Treat.* **154**, 13–22 (2015).
154. Garrido-Laguna, I. *et al.* Tumor engraftment in nude mice and enrichment in stroma- related gene pathways predict poor survival and resistance to gemcitabine in patients with pancreatic cancer. *Clin. Cancer Res.* **17**, 5793–5800 (2011).
155. Delitto, D. *et al.* Patient-derived xenograft models for pancreatic adenocarcinoma demonstrate retention of tumor morphology through incorporation of murine stromal elements. *Am. J. Pathol.* **185**, 1297–1303 (2015).

156. Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
157. Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
158. Dunne, P. D. *et al.* Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clin. Cancer Res.* **22**, 4095–4104 (2016).
159. Bhargava, M. *et al.* Scatter factor and hepatocyte growth factor: activities, properties, and mechanism. *Cell Growth Differ.* **3**, 11–20 (1992).
160. Pennacchietti, S. *et al.* Microenvironment-derived HGF overcomes genetically determined sensitivity to anti-MET drugs. *Cancer Res.* **74**, 6598–6609 (2014).
161. Mestas, J. & Hughes, C. C. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
162. Brodeur, J. *et al.* Knock-in of human HGF into the mouse genome maintains endogenous HGF regulation and supports the growth of HGF-dependent human cancer cell lines. *Cancer Res.* **69**, abstr. 305 (2009).
163. Zimmer, L. *et al.* Phase I expansion and pharmacodynamic study of the oral MEK inhibitor RO4987655 (CH4987655) in selected patients with advanced cancer with RAS–RAF mutations. *Clin. Cancer Res.* **20**, 4251–4261 (2014).
164. Eckhardt, S. G. *et al.* Challenges, opportunities, and lessons learned in the bench-to-bedside translation of xenopatient studies. *Clin. Cancer Res.* **22** (16 Suppl.), abstr. IA20 (2016).
165. Townsend, E. C. *et al.* The public repository of xenografts enables discovery and randomized phase II-like trials in mice. *Cancer Cell* **29**, 574–586 (2016).
166. Baralis, E., Bertotti, A., Fiori, A. & Grand, A. LAS: a software platform to support oncological data management. *J. Med. Syst.* **36** (Suppl. 1), S81–S90 (2012).
167. Chou, J. *et al.* Phenotypic and transcriptional fidelity of patient-derived colon cancer xenografts in immune-deficient mice. *PLoS ONE* **8**, e79874 (2013).
168. Ito, R., Takahashi, T., Katano, I. & Ito, M. Current advances in humanized mouse models. *Cell. Mol. Immunol.* **9**, 208–214 (2012).
169. Conway, T. *et al.* Xenome — a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, i172–i178 (2012).
170. Bacac, M. *et al.* A novel carcinoembryonic antigen T-cell bispecific antibody (CEA TCB) for the treatment of solid tumors. *Clin. Cancer Res.* **22**, 3286–3297 (2016).
171. Ito, M. *et al.* NOD/SCID/γ_c^{null} mouse: an excellent recipient mouse model for engraftment of human cells. *Blood* **100**, 3175–3182 (2002).
172. Shultz, L. D. *et al.* Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2Rγ^{null} mice engrafted with mobilized human hemopoietic stem cells. *J. Immunol.* **174**, 6477–6489 (2005).
173. Shultz, L. D., Brehm, M. A., Garcia-Martinez, J. V. & Greiner, D. L. Humanized mice for immune system investigation: progress, promise and challenges. *Nat. Rev. Immunol.* **12**, 786–798 (2012).
174. Traggiai, E. *et al.* Development of a human adaptive immune system in cord blood cell-transplanted mice. *Science* **304**, 104–107 (2004).
175. Ito, R. *et al.* Establishment of a human allergy model using human IL-3/GM-CSF-transgenic NOG mice. *J. Immunol.* **191**, 2890–2899 (2013).
176. Billerbeck, E. *et al.* Development of human CD4$^+$FoxP3$^+$ regulatory T cells in human stem cell factor-, granulocyte-macrophage colony-stimulating factor-, and interleukin-3-expressing NOD-SCID IL2Rγ^{null} humanized mice. *Blood* **117**, 3076–3086 (2011).
177. Rongvaux, A. *et al.* Development and function of human innate immune cells in a humanized mouse model. *Nat. Biotechnol.* **32**, 364–372 (2014).
178. Cunningham, D. *et al.* Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N. Engl. J. Med.* **351**, 337–345 (2004).
179. Kawazoe, A. *et al.* A retrospective observational study of clinicopathological features of KRAS, NRAS, BRAF and PIK3CA mutations in Japanese patients with metastatic colorectal cancer. *BMC Cancer* **15**, 258 (2015).

Acknowledgements

The authors would like to thank all members of the EurOPDX Consortium who also contributed to this article, and in particular S. Corso, S. Giordano, P. P. López-Casas, K. Moran-Jones and F. Nemati. The Caldas laboratory would like to thank the PGE team for their support, especially Lisa, Steve and Yi. A.T.B. is supported by Science Foundation Ireland under grants 13/CDA/2183 and 15/TIDA/2963 and further receives funding from the Irish Cancer Society Collaborative Cancer Research Centre BREAST-PREDICT Grant CCRC13GAL. D.G.A. and R.B.C. are supported by Breast Cancer Now. F.A., E.H. and J.C.M. received KULeuven GOA funding (GOA/14/012) and a research grant from Stichting tegen Kanker. J.A. is funded by the Breast Cancer Research Foundation, the Spanish Association Against Cancer (AECC) and the Instituto de Salud Carlos III (PI16/00253 and CIBER-ONC). A.V.B. and D.K.C. are supported by Cancer Research UK (C29717/A17263), the Wellcome Trust (10372/Z/14/Z), the Scottish Genomes Partnership — SEHHD-CSO 1175759/2158447, the Howat Foundation and Pancreatic Cancer UK. A.B., C.C. and O.M.R. have been supported by funding from Cancer Research UK and by the European Union to the EUROCAN Network of Excellence (FP7; grant number 260791). E.B. is supported by the CETOCOEN PLUS project (CZ.02.1.01/0.0/0.0/15_003/0000469) and the RECETOX Research Infrastructure (LM2015051). G.C. and D.V. were funded by NIH transformative R01CA156695 and European Research Council (ERC) Advanced grant 1400206AdG-322875. S.G.E. receives support from NCI grant 1UM1CA186688 for early-phase trials through the ET-CTN. E.G.S. is supported by the Spanish Ministry of Economy and Competitivity MINECO and from the ISCIII (SAF2014-55997; PIE13/00022, co-funded by FEDER funds/European Regional Development Fund (ERDF) — a way to build Europe), by a Career Catalyst Grant from the Susan Komen Foundation (CCR13262449) and by a European Research Council Consolidator grant (CoG682935). M.A.J. is supported by an Irish Health Research Board Health Research Award (#HRA-POR-2014-547). S.D.J. is supported by the Dutch Cancer Society (grants RUG 2010-4833, RUG 2011-5231, RUG 2012-5477 and RUG 2014-6691). J.J. is funded by the Dutch Cancer Society (NKI 2011-5197 and EMCR 2014-7048), the Netherlands Organisation for Scientific Research (Zenith 93512009, Vici 91814643, CancerGenomiCs.nl) and the European Research Council (ERC-SyG CombatCancer). K.K. and D.S.P. are supported by the Dutch Cancer Society (NKI-2013-5799). L.L. and P.G.P. are funded by ERC Advanced Grant 341131 and Italian Association for Cancer Research (AIRC) Investigator Investigator Grant 14216. G.M.M. receives funds from the Norwegian Cancer Society (421851) and the Research Council of Norway (222262/F20). J.H.N. is funded by the Research Council of Norway under grant 250459/F20. H.G.P. is supported by the Instituto de Salud Carlos III and the Miguel Servet Program (MSII14/00037). V.S. is supported by the Instituto de Salud Carlos III (PI13/01714 and the Miguel Servet Program CP14/00028), by a Career Catalyst Grant from the Susan Komen Foundation CCR15330331 and the FERO Foundation. L.S. was funded by Worldwide Cancer Research (WCR/AICR Grant #13-1182), the European Research Council (CoG Grant #617473), the Instituto de Salud Carlos III (FIS Grant #PI13/01705) and the FERO Foundation. A.V. is supported by the Instituto de Salud Carlos III (PI13/0133 and PIE13/00022 (Oncoprofile)), Fundación Mutua Madrileña AP150932014 and a grant from the Spanish Association Against Cancer from Barcelona, AECC. A.B. is supported by AIRC (Investigator Grant project 15571). L.T. and E.M. are supported by the AIRC (Special Programme Molecular Clinical Oncology 5 × 1000, project 9970, and Investigator Grant projects, 14205 to L.T. and 12944 to E.M.) and also receive funding from the Fondazione Piemontese per la Ricerca sul Cancro-ONLUS (5 × 1000 Italian Ministry of Health 2011).

Competing interests statement

The authors declare competing interests: see Web version for details.

DATABASES

Children's Oncology Group (COG) cell culture and xenograft repository: http://www.cogcell.org/xenografts.php
Public Repository of Xenografts (PRoXe): http://www.proxe.org
US National Cancer Institute (NCI) repository of patient-derived models: https://dtp.cancer.gov/repositories/

FURTHER INFORMATION

EurOPDX: http://www.europdx.eu
GENiE: http://worm-genie.eu/
TRANSAUTOPHAGY: http://cost-transautophagy.eu/
US Pediatric Preclinical Testing Consortium (PPTC): http://www.ncipptc.org/

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

[***12***] Popovici V, **Budinská E**, Čápková L, Schwarz D, Dušek L, Feit J, Jaggi R. Joint analysis of histopathology image features and gene expression in breast cancer. BMC Bioinformatics. 2016 May 11;17(1):209. doi: 10.1186/s12859-016-1072-z. PMID: 27170365; PMCID: PMC4864935.

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Joint analysis of histopathology image features and gene expression in breast cancer

Vlad Popovici[1*], Eva Budinská[1,2], Lenka Čápková[1], Daniel Schwarz[1], Ladislav Dušek[1], Josef Feit[1] and Rolf Jaggi[3]

Abstract

Background: Genomics and proteomics are nowadays the dominant techniques for novel biomarker discovery. However, histopathology images contain a wealth of information related to the tumor histology, morphology and tumor-host interactions that is not accessible through these techniques. Thus, integrating the histopathology images in the biomarker discovery workflow could potentially lead to the identification of new image-based biomarkers and the refinement or even replacement of the existing genomic and proteomic signatures. However, extracting meaningful and robust image features to be mined jointly with genomic (and clinical, etc.) data represents a real challenge due to the complexity of the images.

Results: We developed a framework for integrating the histopathology images in the biomarker discovery workflow based on the bag-of-features approach – a method that has the advantage of being assumption-free and data-driven. The images were reduced to a set of salient patterns and additional measurements of their spatial distribution, with the resulting features being directly used in a standard biomarker discovery application. We demonstrated this framework in a search for prognostic biomarkers in breast cancer which resulted in the identification of several prognostic image features and a promising multimodal (imaging and genomic) prognostic signature. The source code for the image analysis procedures is freely available.

Conclusions: The framework proposed allows for a joint analysis of images and gene expression data. Its application to a set of breast cancer cases resulted in image-based and combined (image and genomic) prognostic scores for relapse-free survival.

Keywords: Histopathology images, Image analysis, Biomarker discovery, Gene expression, Multimodal data mining

Background

The recent technological progress made scanning the whole pathology slides affordable and its integration in the routine pathology workflow feasible. This resulted in a revived interest in developing new computational methods for nuclear morphometry and tissue architecture characterization, as well as for developing new tissue-based biomarkers [1]. In the last decade, genomic and proteomic techniques have been the methods of choice for novel biomarker discovery. When applied to the same sample, the pathology imaging and *omics technologies allow the investigation of the underlying biology from different perspectives, increasing the chances for identifying effective biomarkers. Ideally, these perspectives could be integrated in a common data analytical framework, to enable a joint (or multimodal) data mining and decision [2].

Traditionally, the methods for analyzing pathology images focused on extracting quantitative measures for a set of predefined morphological parameters (e.g. counting, classifying and characterizing the nuclei) and on reproducing the expert's decision in diagnostic applications (for a review see Gurcan et al. [3]). More recently, a number of applications of pathology image analysis combined image-based quantitative features with genomic

*Correspondence: popovici@iba.muni.cz
[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic
Full list of author information is available at the end of the article

information. For example, Yuan et al. [4] showed that nuclear morphometry is an independent prognostic factor that can improve a genomic signature. A similar approach is discussed by Kong et al. [5] in the case of glioblastoma where they show how nuclear and cytoplasmic features can be linked to genomic profiles and survival outcome. More advanced techniques combine several image-derived characteristics, such as co-localization of tumor nuclei and lymphocyte infiltration [6]. In all these cases however, the imaging features were predefined and based on previous known associations between histopathology and diagnostic/prognostic.

Our interest is in developing a more general computational framework that would allow the integration of the standard histopathology images in the biomarker discovery workflow and in which the image features would be learned in a data-driven fashion, enabling a prior-free data mining. The main challenge when analyzing the pathology images stems from their high complexity and size, and seeming incompatibility with *omics data. In the present work we propose to use the *bag-of-features* approach [7] for reducing the dimensionality of the images and extracting salient features. This approach has already been used in histopathology image classification applications [8, 9] and has the main advantage of allowing an unsupervised learning of image representation. The features extracted describe mostly the textural appearance of small neighborhoods and may be combined with other types of features (e.g. nuclear morphometry) in later stages of image analysis, but these approaches will not be discussed here. As an alternative to bag-of-features, one could use deep learning methods for learning image features as proposed by Cireşan et al. [10] or Cruz-Roa et al. [11]. However, these methods require a larger sample size and were applied in a supervised learning context.

We propose a novel representation of histopathology images which extends the standard bag-of-features with a number of derived measurements aimed at capturing more global characteristics of the tissue sample. In addition, we introduce an objective criterion for optimizing the image representation. The new computational framework is demonstrated in a biomarker discovery scenario, where prognostic features (both imaging and gene expression) for relapse-free survival in breast cancer are sought. We see the application of this approach as a succession of two independent steps, not necessarily performed on the same data corpus. In the first step, a histopathology image representation is learned from a collection of images representative for the pathology under investigation. In the second step, the images of interest are recoded based on the constructed representation and the resulting image features are jointly analyzed with the molecular and clinical data.

Methods

Data

The data used in this study is a subset of the data from Moor et al. [12], selected solely based on the availability of the material for analysis. Overall there were $n = 196$ standard pathology (haematoxylin-eosin-stained) slides with breast tissue sections, not all containing a tumoral component and not necessarily from different cases. All images were obtained by whole-slide scanning of the pathology slides at 40× magnification, resulting in color images of about $150,000 \times 100,000$ pixels.

These data were partitioned into an image model learning set ($n = 131$) and a biomarker discovery/data mining set ($n = 65$). In the biomarker discovery set we kept unique cases for which the slides contained $> 70\ \%$ tumor component and the clinical, survival and gene expression data were all available. The expression profiles of 47 target genes (including 5 control genes) were obtained by quantitative real-time PCR (qRT-PCR). A full description of the data set is available in Moor et al. [12] and the major characteristics of the biomarker discovery set used here are given in Additional file 1.

We computed the genomic prognostic signature (PRO_10) as described in Antonov et al. [13] for all the cases with full genomic profiles.

Image processing

Preprocessing

All images were downscaled to an equivalent of 2.5× magnification by subsampling the Gaussian-filtered higher resolution images (the 4-th level in a Gaussian pyramid). In the resulting images a mask corresponding to the tissue regions was obtained by adaptive thresholding in the green channel. The mask was subsequently refined by morphological operations: erosion with a circular structuring element with radius 13 followed by gap filling and removal of small objects.

For each image we estimated the intensity of haematoxylin (H) staining by deconvolving the RGB-images as described by Ruifrok et al. [14]. The intensity levels of the haematoxylin image (H-image) were adjusted by adaptive histogram equalization. Finally, the background pixels were masked out using the tissue region mask computed as above. In all subsequent image processing steps, only the H-images were used.

Learning the image representation

The bag-of-features [7] approach has two main stages: (i) learning an appropriate *codebook* for representing the images of interest and (ii) re-coding the images based on the frequencies of each *codeblock* (codeword from the codebook). Thus, the resulting representation of the image is a histogram of the codeblocks. For the current application, we extended this representation to include several derived features. We point out that once an appropriate image representation is learned, it can be applied unchanged to other similar image collections thus this step does not need to be repeated on each new data set.

Codebook learning The codebook is a collection of representative local descriptors $\{C_1, \ldots, C_K\}$ obtained as centers of K clusters resulting from k-means clustering of a number of image local descriptors (i.e. a vector quantization procedure). For this, the images are decomposed in a set of local neighborhoods for which descriptor vectors are computed. The local descriptors range from pixels intensities to responses to filter banks or other textural descriptor. For the histopathology images, the Gabor wavelets provide a good set of descriptors, so they were adopted in the present work. Each local neighborhood of size $w \times w$ was convolved with a bank of 24 Gabor filters [15],

$$G(x, y; \nu, \theta, \sigma) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \times \exp\left(2\pi \nu j (x \cos\theta + y \sin\theta)\right)$$

where $j = \sqrt{-1}$, ν was the frequency, θ the orientation and σ the bandwidth of the Gaussian kernel. These parameters were set to $\sigma \in \{1, 2\sqrt{2}\}, \theta \in \{k\frac{\pi}{4} | k = 0, \ldots, 3\}$ and $\nu \in \{3/4, 3/8, 3/16\}$, respectively. They were kept fixed throughout all the experiments. For each filter response, its mean and standard deviations were recorded, thus each local neighborhood $w \times w$ was represented by 48 values (24 means and 24 standard deviations). A comparison of Gabor wavelets with other local descriptors, in the context of histopathology image analysis, is given by Budinská et al. [9].

The size of the codebook (i.e. the number of clusters in k-means clustering), K, is a free parameter that has to be chosen/guessed at the moment of codebook construction [8]. It can also be optimized for the problem at hand [9] using, for example, the Gap statistic [16]. Here we took advantage of having available a number of examples for different tissue components (fat, fat foamy macrophages, comedo necrosis, connective tissue and carcinoma infiltrating fat – for examples see Additional file 1) which we used as reference categories. The goal was to choose the size of the dictionary K in such a way that the representations of these categories are sparse and have a minimal overlap. For each image i, let $y_i = \{j \mid \text{if codeblock } C_j$ is used in coding the sample $i\}$, be the set of codeblocks used in its coding. Then we define the following quantities (where $| \cdot |$ denotes the cardinality of a set):

- total Jaccard index,

  $$J(K) = 0.5 \sum \frac{|y_i \cap y_j|}{|y_i \cup y_j|},$$

  where the sum is taken over all pairs (i, j) of images from different reference categories;
- total sum of within-cluster distances,

  $$D(K) = \sum_{k=1}^{K} \sum_{i \in \text{cluster } k} \|\mathbf{x}_i - C_k\|^2,$$

  where $\mathbf{x}_i$ are the descriptor vectors.

With these quantities, we defined an (empirical) objective function:

$$\Psi(K) = \log \frac{n_c(n_c - 1)}{2} - \log J(K) - \log \sqrt{D(K)} - 0.75 \log K,$$

where n_c is the number of reference categories (in our case $n_c = 5$). The overall goal of our image recoding step is to find a low dimensional (sparse) representation which still bears enough information for discriminating major tissue components. For this, we minimize $J(K)$, i.e. the overlap between the representations of the reference categories. At the same time, we require tight clusters (small within-cluster total distances $D(K)$) and sparse representation (small K). Hence, the desired value for K is the one that maximizes $\Psi(K)$, where we note that the first term is constant (included to bring the values closer to 0) and that the scaling factor 0.75 is used to reduce the influence of K.

Image recoding Once a suitable K is found and a codebook is constructed by k-means clustering, the standard bag-of-feature approach represents the images as codeblock histograms. However, in this coding, all spatial information about the distribution of the codeblocks is lost. Consider the situation in Fig. 1a: all four images have the same number of patches assigned to the same codeblock, but the spatial arrangement is very different. In order to characterize these spatial differences, we extend the image representation with a number of statistics on the distribution of the codeblocks. For a given image and for each codeblock $k \in \{1, \ldots, K\}$, we construct a binary image in which 1s represent regions assigned to the codeblock and 0s everything else. In these binary images, the connected components (4-neighbor connectivity) define individual objects and for each of them we compute the area (in pixels) and the compactness index (ratio of the squared perimeter to the area of the object). Finally, for each image and each codeblock, we compute (i) the median area, (ii) the maximum area, (iii) the ratio of the maximum area to the total area of the objects, (iv) the skewness of the distribution of the area values and (v) the mean compactness. Thus, for each codeblock in an image, aside from its frequency, we add five new values aimed at characterizing the distribution of the codeblock in the image. We will refer to these additional quantities as the “extended set of features”. The final representation of an

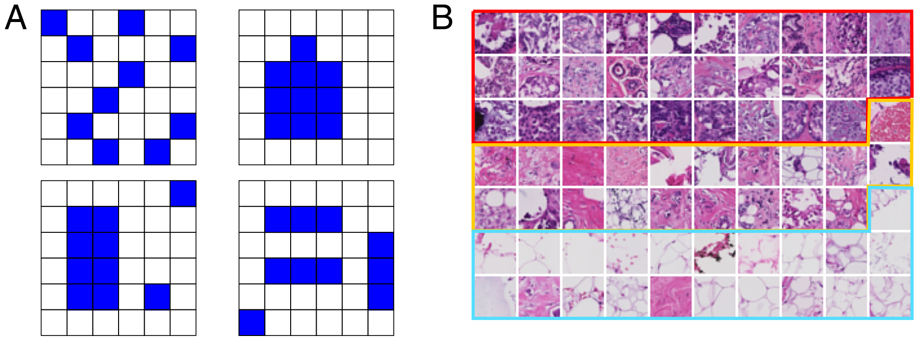


Fig. 1 Codeblocks and codebook. **a** An example of four different hypothetical distributions of the codeblocks leading to identical frequencies. To cope with such situations, the distribution of codeblocks is also taken into account through extended image features. **b** A visual representation of the obtained codebook. The 70 image patches are the closest to the codeblocks obtained after *k*-means clustering. The three groups of codeblocks (with 29, 20 and 21 elements, respectively) correspond to the major clusters in Fig. 2 and the ordering of the image patches is the same as in the clustering

image has a length of $6K$: K values for the codeblock histogram (the standard representation) and $5K$ values of the extended representation.

Joint data mining

The new representation of the images allows for direct application of standard data mining techniques. In the case of multi-modality data mining, the choice of a proper similarity metric/measure is of crucial importance. Two main strategies may be attempted for defining a proper similarity: combination of single, modality-specific, metrics or building/learning a fully multi-modality metric. The first approach has the advantage of using established metrics usually resulting in easily interpretable models and facilitating the comparison with known results. The second approach promises to build a similarity metric that better exploits the multi-modality nature of the data. These ideas can be implemented, for example, in the context of kernel machines (such as Support Vector Machines) where composite kernels (based on closure properties – see [17] p.75) would represent a possible implementation of the first approach and multiple kernel learning [18] an implementation of the latter.

In the present work and in order to demonstrate the general analytical framework, we make use of standard statistical tools. We aim at identifying image features that could be linked to expression levels of the genes of interest (genotype-phenotype association) and potential image biomarkers that alone or in combination with gene expression can be used for defining a prognostic signature. Besides the gene expression, we also used a proliferation gene signature PRO_10 [12, 13], which was shown to be prognostic in various cohorts of patients with breast cancer.

To test the association between image features and tumor size (T) and grade (G) we dichotomized the clinical variables (T: {T1, T2} vs {T3, T4}, and G: {G1,G2} vs. G3, respectively) and used two-sided t-test, with 0.05 significance level. The association of image features with gene expression was assessed based on correlation test (Pearson) with significance level 0.05 and the condition that the correlation coefficient was at least 0.5 (in absolute value). We also used canonical correlation analysis (CCA) to study the associations between image features and molecular data with significance level of 0.05 for Wilks' test. The association between image features and survival outcome (relapse-free survival – RFS) was tested using Cox proportional hazard models (log-likelihood test), with significance level of 0.05. The hazard ratios were estimated from interquartile range-standardized variables (both image and genomic variables). To test if an image feature improves the prognostic value of the gene signature, we tested the difference between the models with and without the variable of interest using likelihood ratio tests. To assess the difference in survival between two groups we used log-rank tests. We binarized the variables by their median value, into high- and low- expressions or values. Since the work reported here is purely exploratory and the sample size is rather small, no adjustment for multiple hypotheses testing was performed. We used hierarchical clustering (Ward method) with Euclidean distance between samples to cluster the codeblocks.

All statistical analyses were performed in `R` package for statistical computing (http://www.r-project.org) version 3.2.2.

Results

Codebook

The image analysis methods described above were implemented in a `Python` package (available at https://github.com/vladpopovici/WSItk), using the `scikit-image` [19] and `Mahotas` [20] libraries.

For the codebook construction we used only the modeling set of images, none of the image used in the data mining phase being used for learning the codebook. From each image, a set of 3000 random patches of size 32 × 32 was extracted and the corresponding Gabor descriptors computed (vectors of 48 elements). These descriptor vectors were clustered using the k-means algorithm to build the codebooks. We estimated the optimal (in the sense of the Ψ objective function, described above) codebook size by evaluating $\Psi(k)$ for $k = 10, 20, \ldots, 1000$. The optimal value was found to be $K = 70$ (see Additional file 1 for a plot of $\Psi(k)$) leading to 420 feature vectors for each image. Since the codeblocks are centers of the clusters (the means of descriptor vectors assigned to the respective cluster), they might not necessarily correspond to observed image regions. Thus we selected the closest regions to the codeblocks (the corresponding descriptor vectors were the closest to the codeblocks) to provide an approximate visual representation of the codebook - Fig. 1b. In the following, to designate a specific codeblock from the codebook, we will use the notation *C.xy*. We have extensively investigated the stability of the learned codebooks and the resulting image representations and we found the process to be stable – see Additional file 1.

The hierarchical clustering of the codeblocks (Fig. 2) revealed a rather structured content: three major groups of codeblocks could be identified. We tentatively labeled them as “proliferation patterns”, “invasion/differentiation patterns/connective tissue” and “sparse tumor nuclei/differentiation/fat” to indicate the major components in the clusters - without claiming a precise histopathological characterization.

A number of codeblocks were found to be associated with tumor size (C.10, C.18, C.29, C.38, C.41, and C.42) and grade (C.09, C.34, C.43, C.45, C.48, and C.62).

Correlations between image features and gene expression

The association analysis between image features and gene expression identified a number of significant ($p < 0.05$ and $\rho > 0.5$) pairwise correlations (all in the range $0.50 - 0.60$). In all, eight different codeblocks were associated with different genes, most of them with *CCNE1* and *CCNB2*. The codeblock C.31 was associated with most genes (*CCNE1, CCNB2, BIRC5, PRC1, SPAG5*) either by its frequency of appearance in the image or by the skewness of its distribution. By summing the frequencies corresponding to image features that are highly correlated (e.g. C.38, C.31, C.01, C.51, C.41, C.68) the correlations coefficients were improved to $0.65 - 0.70$. CCA confirmed the association between these image features and gene expression data (Wilks’ test $p = 0.026$). The image features C.10, C.19, C.57, and C.68 and the genes *CCNE1*, *CCNB2*, and *SPAG5* had the strongest impact on the canonical dimensions. These were also the most stable image features-gene expression correlations in the image representation stability experiments – see Additional file 1.

Despite the fact that the PRO_10 gene signature is an average of proliferation genes which were found to be

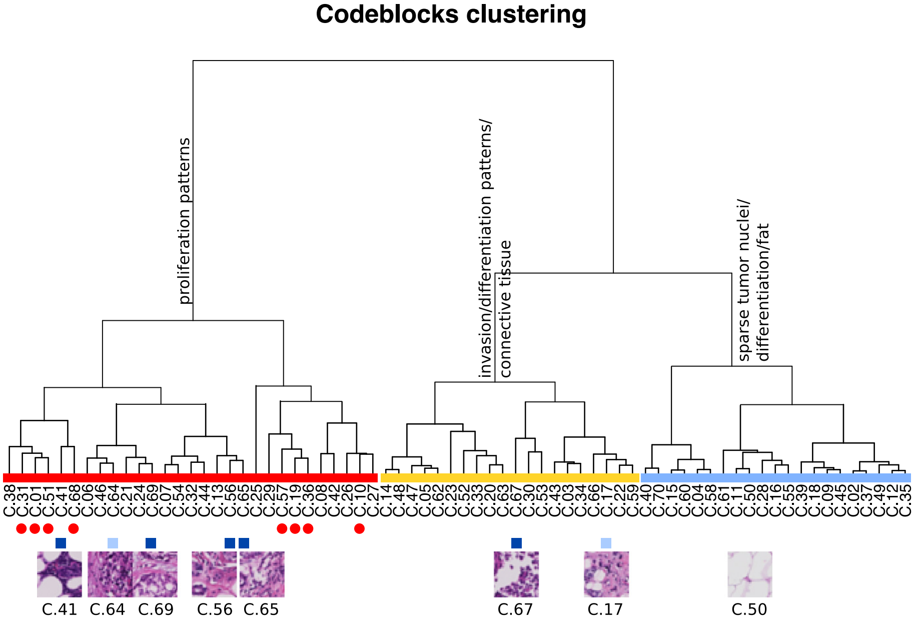


Fig. 2 Hierarchical clustering of the codebook. Clustering the codeblocks led to identification of three major clusters, to which generic terms have been assigned. The codeblocks correlated with gene expression are marked with *red dots*. The codeblocks with potential prognostic value (in univariate analysis) are marked with blue squares (*dark blue* for p-value < 0.01, *light blue* for $0.01 \leq p$-value ≤ 0.05

correlated with image features, the correlations between image features and PRO_10 did not reach the required significance level in all but one case: the skewness of codeblock C.31.

Survival analyses

The goal of the analyses performed was to assess the utility of image-based variables for predicting relapse-free survival independently, or combined with the PRO_10 signature. In the set of samples analyzed, the genomic score is a strong prognostic marker (Cox regression: $p = 0.001, \mathrm{HR} = 2.12, 95\ \%\ \mathrm{CI} = (1.29, 3.51)$).

Univariate Cox proportional hazards models were fit for each of the 420 image features resulting in the identification of several significant associations with relapse-free survival endpoint. The most prognostic image features were C.41, C.56, C.65, C.67, C.69, with $p < 0.01$ and HR between 1.16 and 1.70. From the extended set of features, the median area of the regions assigned to clusters C.15 and C.26 were significantly associated with RFS ($p <$ 0.05). The strongest predictor among the image features was C.69 ($p = 0.0018, \mathrm{HR} = 1.7, 95\ \%\ \mathrm{CI} = (1.22, 2.37)$).

In combined models (image feature and genomic score) a number of image features led to improved models (likelihood ratio test $p < 0.05$), most of them from the extended set of features. From all these image features, C.69 remained significant in the multivariate model (with PRO_10) and had no significant interaction with the genomic signature.

We defined an image score variable by averaging C.41, C.56, C.65, C.67, C.69 which resulted in a stronger prognostic factor (Cox regression: $p = 0.0003$ and $\mathrm{HR} = 1.76, 95\ \%\ \mathrm{CI} = (1.30, 2.40)$ - see also Figure 3). In a regression model including the genomic and the image scores, both remained independent significant variables (PRO_10: $p = 0.05$, image score: $p = 0.007$, no significant interaction) and the model was signficantly better than the corresponding univariate models ($p = 0.013$). In Fig. 4 the Kaplan-Meier curves for binarized (by median value) scores are shown, together with corresponding p-values (log-rank tests) and hazard ratios. Another visualization of the prognostic scores is given in Fig. 5 where the expected survival at 4 years is shown as a function of the genomic, image-based, and combined scores, respectively. Two examples of high risk cases, according to the image-based score, are given in Additional files 2 and 3.

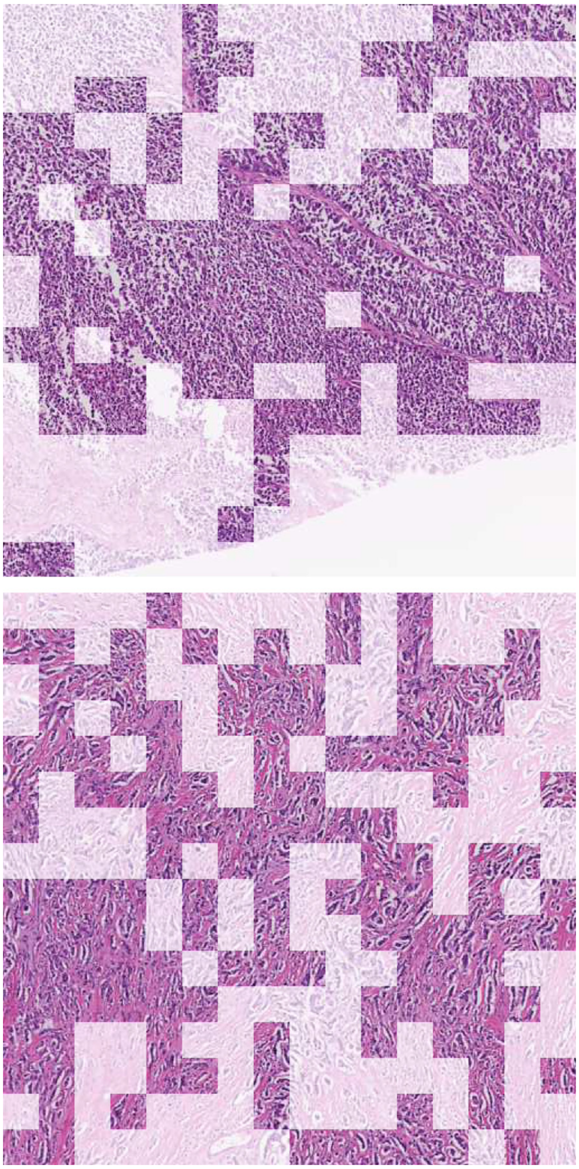


Fig. 3 Regions assigned to the most prognostic codeblocks. 512 × 512 regions from two different samples with high image score (high risk of relapse), at 2.5× magnification. The image patches represented in full color were assigned to one of the C.41, C.56, C.65, C.67 or C.69 codeblocks. In Additional files 2 and 3, the corresponding whole slide images are provided

Discussion

The main challenge in introducing the histopathology images in the general data mining biomarker discovery framework stems from their high complexity and low level of information representation. Thus, while the images contain a huge amount of data (in the order of 10^{10} pixels) the extraction of information implies a considerable effort. Traditionally, this effort is performed by the expert pathologists or, more recently, by using quantitative methods for measuring a set of predefined morphological aspects to complement the pathology report. In this work, we took a third approach, in which the image data is reduced to a number of essential patterns (the codeblocks) whose frequency and spatial distribution in the image is used for data mining. The codeblocks are learned independent of any prior knowledge about the images, potentially

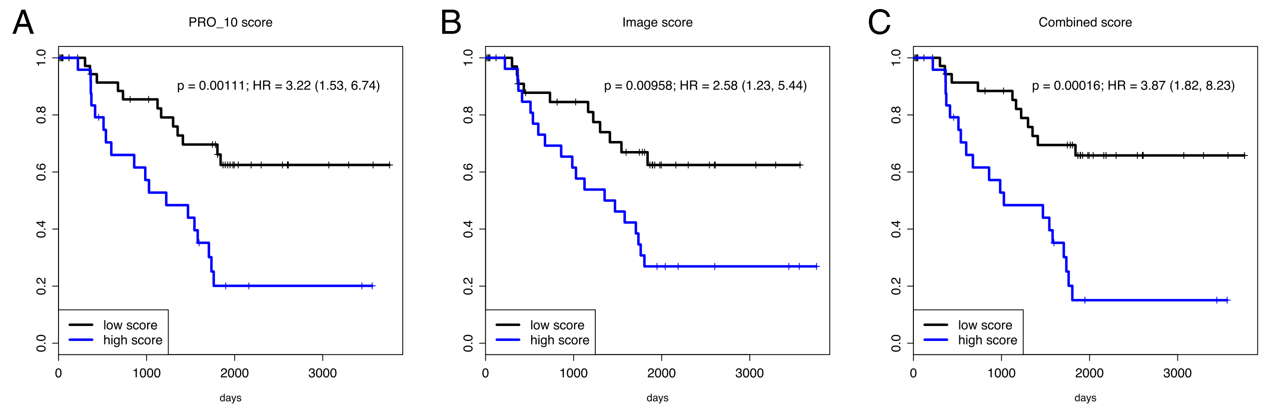


Fig. 4 Kaplan-Meier curves for binarized scores. The genomic (**a**), image-based (**b**) and combined scores (**c**) were binarized by the respective median values into "low score" (low risk) and "high score" (high risk) categories. The combined score slightly improves on the genomic score

enabling the discovery of new image features not necessarily assessed during the pathology review of the cases. The obvious drawback is the difficulty of interpreting some of the patterns and the possibility of having also artifacts in the model. The adopted representation of local neighborhoods in the image (responses to a bank of Gabor filters) encouraged the identification of codeblocks with distinctive textural appearance (Fig. 1). This local appearance may be later on combined with a nuclei detector and classifier (as in Yuan et al. [4]), for example, to obtain a more comprehensive characterization of the image.

By examining the similarities between codeblocks, we identified three major aspects of the images that are captured: proliferation, invasion/differentiation (within connective tissue) and isolated tumor nuclei (within regions predominantly with fat component) (Fig. 2). This result combined with the observation that the whole third cluster did not contribute to the prognostic models, suggests a possible refinement of the current method, in which these regions with high fat content are discarded in an initial preprocessing stage and a more detailed model is used to characterize the remaining regions.

We demonstrated the integration of the image features in a standard biomarker discovery scenario, in which both image-genes correlations (precursors to genotype-phenotype associations) as well as various survival prognostic models were tested. Since the main purpose of this exercise was to demonstrate the integration of image features with genomic information and the sample size was relatively modest, we did not adjust for multiple hypotheses testing and restricted ourselves to an exploratory analysis. Thus the associations found, while hypothesis-generating, have to be taken with caution and more validation is needed.

Most of the genes in the panel were related to proliferation processes, thus it is not surprising that the correlations with image features involved almost exclusively these genes. The strongest associations were found

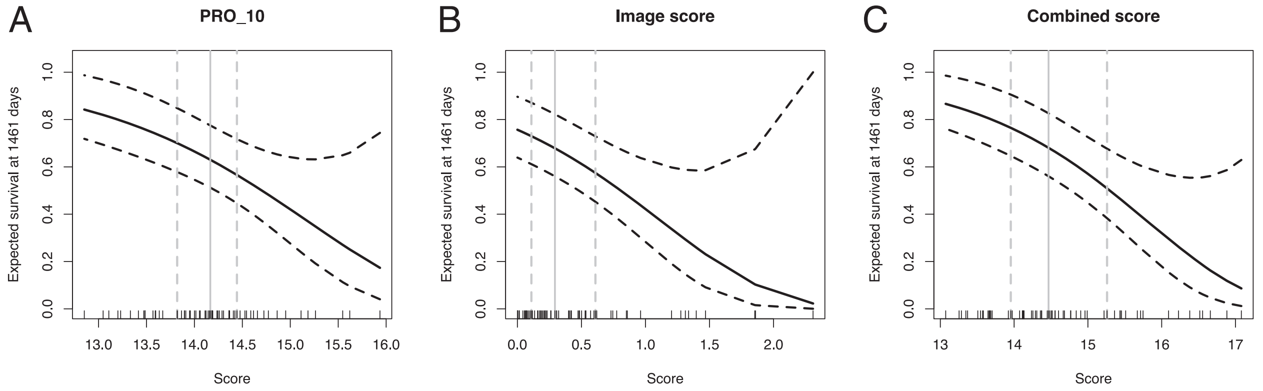


Fig. 5 Prognostic scores at 4 years. Predicting the likelihood of an event (relapse) at 4 years, based on genomic signature (PRO_10 - panel **a**), the image-based score (panel **b**) and the combined score (panel **c**)

with *CCNE1* and *CCNB2*. Somehow surprising, no significant correlation was found with *MKI67* gene, a common marker (with Ki-67 specific staining) for proliferation.

A number of image features were found to be prognostic for RFS and we proposed a simple image-based prognostic score which averages five basic image features. The new score is strongly prognostic and is not correlated with the genomic score considered (PRO_10). When combining the two scores in a multivariable Cox regression, the two remained significant (with a marginal significance for the genomic score) and independent predictors (no significant interaction) leading to an improved model. Thus, the image-based score can be used either alone - as a first line predictor - or in combination with the genomic predictor. These results also demonstrate the complementarity of the two modalities - histopathology imaging and genomics - and suggest that refined predictors can be built by a combination thereof.

It must be noted that the sample size and the number of events did not allow for more variables in the regression models. Further analysis of the scores (either image-based or combined) in the context of usual clinical predictors (TNM-staging, hormonal status, etc.) is required before a definite conclusion about its clinical utility can be drawn. Nevertheless, the image-based score can already be used in applications like searching or indexing in histopathology image archives.

Conclusions

We proposed a general framework for integrating the histopathology images in the routine genomic data analysis pipeline. The image features used are based on the responses of Gabor filters applied to single channel images. The approach can easily be extended to exploit the full color information and to include other types of features.

When applying our method to a data collection of breast cancer samples, we were able to identify a number of associations between image features and gene expression levels. More importantly, several prognostic image features were identified, some of them complementary to the genomic score. Thus, we could build an image-based and a combined survival score, improving on the performance of the genomic score. These results must be validated in larger data sets.

The code implementing the methods described is made freely available and continues to be under active development.

Availability of data and materials

The source code for the image analysis methods described in the paper is available from the GitHub repository https://github.com/vladpopovici/WSItk.

The data used to demonstrate the methods described is not publicly available.

Ethics approval and consent to participate

The data used to demonstrate the methods in this study has been graciously provided by the Department of Medical Oncology, Inselspital Bern, Switzerland. All patients gave a general consent for the use of their tissue samples in research.

Additional files

Additional file 1: Codebook construction details [PDF file]. The codebook was optimized based on a objective function and a set of reference categories. This file contains the plot of the objective function and example images for the selected categories. (PDF 12390 kb)

Additional file 2: High risk carcinoma according to image-based score (Example 1). [JPG file]. Whole-slide image of a tumor labeled as high risk by the image score, with the regions used in scoring highlighted. (JPG 9758 kb)

Additional file 3: High risk carcinoma according to image-based score (Example 2). [JPG file]. Whole-slide image of a tumor labeled as high risk by the image score, with the regions used in scoring highlighted. (JPG 12800 kb)

Abbreviations

HR: hazard ratio; qRT-PCR: quantitative real-time polymerase chain reaction; RFS: relapse-free survival.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VP initiated and designed the study, performed image and statistical data analyses. EB and LB performed bioinformatics and biostatistics analyses. JF performed image analyses and contributed pathology interpretation of the images. RJ performed the gene expression profiling, provided clinical and biological interpretation of the results. DS and LD performed image and statistical analyses. All authors contributed to the manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the support of Ministry of Health of the Czech Republic through grant NT/14134. The results presented here reflect only the authors' views.

Funding

This study was funded by the Ministry of Health of the Czech Republic through grant NT/14134.

Author details

[1]Institute of Biostatistics and Analyses, Faculty of Medicine, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic. [2]RECETOX, Masarykova Univerzita, Kamenice 5, 62500 Brno, Czech Republic. [3]Department of Clinical Research, Faculty of Medicine, University of Bern, Bern, Switzerland.

Received: 9 December 2015 Accepted: 4 May 2016
Published online: 11 May 2016

References

1. Hamilton PW, Bankhead P, Wang Y, Hutchinson R, Kieran D, McArt DG, James J, Salto-Tellez M. Digital pathology and image analysis in tissue biomarker research. Methods. 2014;70(1):59–73.
2. Colen R, Foster I, Gatenby R, Giger ME, Gillies R, Gutman D, Heller M, Jain R, Madabhushi A, Madhavan S, Napel S, Rao A, Saltz J, Tatum J, Verhaak R, Whitman G. NCI Workshop Report: Clinical and Computational

Requirements for Correlating Imaging Phenotypes with Genomics Signatures. Transl Oncol. 2014;7(5):556–69.

3. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng. 2009;2: 147–71.
4. Yuan Y, Failmezger H, Rueda OM, Ali HR, Graf S, Chin SF, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, Johnson N, Doyle S, Turashvili G, Provenzano E, Aparicio S, Caldas C, Markowetz F. Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. Sci Transl Med. 2012;4(157):143.
5. Kong J, Cooper LAD, Wang F, Gutman DA, Gao J, Chisolm C, Sharma A, Pan T, Van Meir EG, Kurc TM, Moreno CS, Saltz JH, Brat DJ. Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. IEEE Trans Biomed Eng. 2011;58(12):3469–74.
6. Nawaz S, Heindl A, Koelble K, Yuan Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. Mod Pathol. 2015;28(6):766–77.
7. Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. Work Stat Learn Comput Vision ECCV. 200459–74.
8. Caicedo JC, Cruz A, Gonzalez FA. Histopathology Image Classification Using Bag of Features and Kernel Functions In: Combi C, Shahar Y, Abu-Hanna A, editors. 12th Conference on Artificial Intelligence in Medicine. Berlin Heidelberg: Springer; 2009. p. 126–35.
9. Budinská E, Čápková L, Schwarz D, Dušek L, Jaggi R, Feit J, Popovici V. Gene expression-guided selection of histopathology image features. In: 15th International Conference on Bioinformatics and Bioengineering. Belgrade: IEEE; 2015. p. 1–6.
10. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Berlin Heidelberg: Springer; 2013. p. 411–8.
11. Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks In: Gurcan MN, Madabhushi A, editors. SPIE Medical Imaging. San Diego, USA: SPIE; 2014. p. 904103.
12. Moor AE, Guevara C, Altermatt HJ, Warth R, Jaggi R, Aebi S. PRO_10 – A new tissue-based prognostic multigene marker in patients with early estrogen receptor-positive breast cancer. Pathobiology. 2011;78(3):140–8.
13. Antonov J, Popovici V, Delorenzi M, Wirapati P, Baltzer A, Oberli A, Thurlimann B, Giobbie-Hurder A, Viale G, Altermatt H, Aebi S, Jaggi R. Molecular risk assessment of BIG 1-98 participants by expression profiling using RNA from archival tissue. BMC Cancer. 2010;10(1):37.
14. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001;23(4):291–9.
15. Daugman JG. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A. 1985.
16. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001.
17. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press; 2004.
18. McFee B, Lanckriet GRG. Learning Multi-modal Similarity. J Mach Learn Res. 2011;12:491–523.
19. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, Scikit-image contributors. scikit-image: image processing in Python. PeerJ. 2014;2:e453.
20. Coelho LP. Mahotas: Open source software for scriptable computer vision. J Open Res Softw. 2013;1(1):e3.

[***13***] Popovici V, **Budinská E**, Dušek L, Kozubek M, Bosman F. Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. Bioinformatics. 2017 Jul 1;33(13):2002-2009. doi: 10.1093/bioinformatics/btx027. PMID: 28158480.

Bioinformatics, 33(13), 2017, 2002–2009
doi: 10.1093/bioinformatics/btx027
Advance Access Publication Date: 31 January 2017
Original Paper

Bioimage informatics

Image-based surrogate biomarkers for molecular subtypes of colorectal cancer

Vlad Popovici[1],*, Eva Budinská[2], Ladislav Dušek[1], Michal Kozubek[3] and Fred Bosman[4]

[1]Faculty of Medicine, Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic, [2]Faculty of Science, Research Centre for Toxic Compounds in the Environment, Masaryk University, Brno, Czech Republic, [3]Faculty of Informatics, Masaryk University, Brno, Czech Republic and [4]University Institute of Pathology, University of Lausanne, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on September 17, 2016; revised on November 21, 2016; editorial decision on January 15, 2017; accepted on January 17, 2017

Abstract

Motivation: Whole genome expression profiling of large cohorts of different types of cancer led to the identification of distinct molecular subcategories (subtypes) that may partially explain the observed inter-tumoral heterogeneity. This is also the case of colorectal cancer (CRC) where several such categorizations have been proposed. Despite recent developments, the problem of subtype definition and recognition remains open, one of the causes being the intrinsic heterogeneity of each tumor, which is difficult to estimate from gene expression profiles. However, one of the observations of these studies indicates that there may be links between the dominant tumor morphology characteristics and the molecular subtypes. Benefiting from a large collection of CRC samples, comprising both gene expression and histopathology images, we investigated the possibility of building image-based classifiers able to predict the molecular subtypes. We employed deep convolutional neural networks for extracting local descriptors which were then used for constructing a dictionary-based representation of each tumor sample. A set of support vector machine classifiers were trained to solve different binary decision problems, their combined outputs being used to predict one of the five molecular subtypes.

Results: A hierarchical decomposition of the multi-class problem was obtained with an overall accuracy of 0.84 (95%CI=0.79–0.88). The predictions from the image-based classifier showed significant prognostic value similar to their molecular counterparts.

Contact: popovici@iba.muni.cz

Availability and Implementation: Source code used for the image analysis is freely available from https://github.com/higex/qpath.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

The last two decades witnessed fundamental changes in the way we investigate the biology of living organisms, with technological developments fueling major breakthroughs in our understanding of various pathologies and paving the road towards a personalized medicine. Currently, the researchers are armed with a battery of techniques for interrogating the same biological reality at various scales (from sub-cellular to whole population) and from very diverse perspectives (clinical, imaging, genomic, proteomic, etc.) generating high-throughput multimodal data. The bottleneck is now represented by our limited ability to interpret such data in an integrated way (Li *et al.*, 2016) and the need for a more inter-disciplinary approach is epitomized by large scale projects such as The Cancer Genome Atlas (TCGA). In cancer research, one of the main goals it to identify homogeneous groups of patients—i.e. to stratify the patient population—in the hope of finding the common causes and tailored treatments. Traditional stratification of cancer patients is based on histologic and morphologic assessment of the tumor sample and it still defines the golden standard. Lately, various molecular biomarkers have been proposed for the same purpose. The two perspectives are partly overlapping and partly orthogonal, making their integration more challenging. Our present work focusses on translating a gene expression-based cancer patient population stratification into an image-based biomarker, thus trying to bring transcriptomics data into a histopathologic context.

Colorectal cancer (CRC) is the third most frequent cancer worldwide and the second leading cause of cancer mortality in Europe, with metastatic disease accounting for 40–50% of newly diagnosed patients. At the same time, it is a highly heterogeneous disease in terms of prognosis and its response to therapy. Using whole-genome profiling of large data collections, several systems for subcategorization of CRC have been proposed recently (Budinská *et al.*, 2013; De Sousa *et al.*, 2013; Marisa *et al.*, 2013; Sadanandam *et al.*, 2013; Roepman *et al.*, 2013). In general, they relied on clustering the CRC tumors in order to identify patterns of co-regulation of genes that could be indicative of common oncogenic pathways and coherent treatment responses of these tumors. Our own analysis (Budinská *et al.*, 2013) identified five stable tumor clusters (labeled as Subtypes A, B,. . ., E), but also showed that a relatively high proportion of cases remained unaccounted for by this system. A recent effort (Guinney *et al.*, 2015) to harmonize all these discoveries confirmed the presence of four distinct and reproducible subtypes across all studies, labeled CMS1,. . ., CMS4, which match closely our Subtypes A,. . ., D (Guinney *et al.*, 2015). The current golden standard for the identification of the molecular subtype of a given tumor requires the interrogation of a large panel of genes and the application of a genomic classifier. In the analyses reported here, we will use the subtypes as defined in Budinská *et al.* (2013). There are several reasons for this choice: first, since they were derived from the same gene expression data that accompany the images we use, it is hoped that the subtype assignment is less noisy. Second, in Budinská *et al.* (2013), it is noted that an expert pathologist, when presented with the molecular categorization for a set of cases, was able to identify a number of morphological features that were preferentially enriched in one or a few of the subtypes hence, showing preliminary evidence that such connections exist. And third, we are interested in identifying the imaging support for the five previously identified subtypes.

The problem of recognizing the tumor subtype based on imaging data is not new and probably the most studied is the case of breast cancer. For these cancers, five molecular subtypes are currently considered—Luminal A, Luminal B, basal, Her2-enriched and normal-like (Perou *et al.*, 2000)—and surrogate immunohistochemical stains are available (corresponding to hormonal status of ER, PR and Her2 and the invasion marker Ki-67, respectively). Consequently, automatic stain quantification is the strategy of choice for molecular subtype recognition from image data and it was shown to outperform the human expert (Stålhammar *et al.*, 2016). A systematic review of the connections between histological and molecular subtypes in breast cancer is given in Weigelt *et al.* (2010). Other efforts concentrated on the recognition of the high-risk group of triple negative breast cancers on various imaging platforms (Agner *et al.*, 2014; Dogan and Turnbull, 2012). The quantitative image analysis of pathology slides can also serve as a main means for subtype definition. For example, Chang *et al.* (2011) found five subtypes of glioblastoma, one of which being predictive value and correlated with the expression of several genes. Similarly, Lan *et al.* (2015) propose an alternative subtyping of ovarian cancer based on quantitative analysis of tumor microenvironment. A general approach to the identification of disease subtype based on morphologic analysis of pathology slides is described in Cooper *et al.* (2012).

In the case of CRC, Budinská *et al.* (2013) showed that Subtype A had either serrated or papillary architecture, Subtype B represented typical colorectal adenoma with complex tubular architecture, Subtype C was mucinous or solid trabecular, Subtype D was a mixture of desmoplastic and complex tubular architecture, and Subtype E was mixed (see Budinská *et al.*, 2013 for example images). However, these annotations did not lead to a strong classifier.

This observation—that associations can be found between the molecular subtypes and morphological traits of the tumors—constitutes the starting point of our investigations reported here. Our interest is to construct a histopathology image-based classifier able to predict the molecular subtype of a given tumor section without resorting to any other staining but the standard hematoxylin–eosine. This classifier may be seen as a surrogate image biomarker (actually, as we will see, a combination of several biomarkers) for the molecular subtypes and, to the best of our knowledge, it is the first such biomarker to be proposed. This constitutes the main contribution of our work reported here and it represents a largely improved result from our earlier explorations (Budinská *et al.*, 2016). Equally important, our approach does not rely on predefined morphopathological features: the feature selection is guided by the prediction task. This would allow identifying potentially unknown (or overlooked) image features but may also make the interpretation of the models less obvious.

There are many potential application of such a system once established and well tested. First, since it does not require any special laboratory work, it could be easily integrated in the diagnostic workflow to provide hints about the molecular subtype, with no extra costs. It could also be used for sample stratification and selection for retrospective studies, where large collections of samples could easily be filtered for the subtypes of interest without the need of the much more expensive molecular profiling.

Currently, the molecular subtype is established by profiling the expression of a set of genes from the DNA/RNA extracted from the tumoral region of a tissue section and combining their values through a genomic classifier. The whole process involves a number of parameters (from defining the characteristics of the region to be profiled—tumor content, presence/absence of stroma, etc.—to the cut-offs of the classifiers) that are yet to be formalized, thus being error-prone and leading to noisy labels. While we consider the molecular subtypes as the ground truth our image-based classifier is measured against, one has to keep in mind the somehow fuzzy nature of the class definition. These specific settings of our problem make it even more challenging than the more classical applications in the field of digital/computational pathology.

The rest of the paper is structured as follows: the data and the methods used are described in Section 1, followed by the discussion of the results in Section 2 and conclusions in Section 3.

1 Methods

1.1 Data

The present work is based on the data from a subset of the PETACC3 clinical trial (Van Cutsem *et al.*, 2009) samples. The trial compared two treatment regimens (fluorouracil/leucovorin alone or in combination with irinotecan) in CRC and found no differences between the two. The gene expression data for a set of $n = 688$ samples was used (along with other data sets) in the derivation of the molecular subtypes of CRC (Budinská *et al.*, 2013) and is publicly available from ArrayExpress under accession number E-MTAB-990. In (Budinská *et al.*, 2013) the molecular subtypes (denoted A–E) were assigned to a number of $n = 458$ cases, the rest being considered ambiguous (or representing other low-prevalence subtypes) and were labeled as 'outliers'. From those 458 samples, $n = 300$ cases were selected for this study based purely on technical considerations (availability of histopathology tumor section, acceptable whole slide image quality, tissue sample not too fragmented, etc.). The 'outlier' (from a molecular subtype perspective) cases were not considered in the present study.

All molecular subtypes were represented in this collection with the following frequencies: (A) 21, (B) 140, (C) 37, (D) 81 and (E) 21. The slides were annotated by an expert pathologist and these annotations were present in the digital versions—a typical example is given in Figure 1 (note the annotations delineating the loosely the tumoral and normal tissue components).

From the whole collection of 300 images, a subset of 100 images was selected by stratified random sampling to form the *development set*. This development set was used for selecting the image representation model and, for designing, the classification approach. We did not use the whole available data in order to reduce the likelihood of obtaining a model too adapted to our particular collection of samples (overfitted). For the same reason, we also preferred limiting the number of experiments, comparing only several modeling approaches. The remaining 200 images were added at a later stage when the multi-class classifier performance was estimated by cross-validation. Other strategies of selecting a development set (eventually larger, equal number of cases per class, etc.) could have been attempted, with their own advantages and drawbacks, but we found the chosen approach to provide a reasonable trade-off.

1.2 Image acquisition and preprocessing

All whole slide images of hematoxylin–eosin stained tumor sections were acquired at 20× magnification, using a Hamamatsu NanoZoomer C9600 scanner. The resulting images were compressed by the image acquisition software using JPEG standard (at 80% quality) and stored in the proprietary NDPI format. The resolution of the images was 455nm/pixel (equivalent of 55824 DPI) for a typical size of $100\,000 \times 50\,000$ pixels (depending on the size of the tissue section). The images were exported in standard TIFF format using OpenSlide software library (Satyanarayanan *et al.*, 2013).

The images were down-scaled to an equivalent 10× magnification and only tumoral regions were retained from each sample (manually cut following the pathologist's annotations)—the pixels outside the tumors being set to zero. For example, the image in Figure 1 contains two tumoral regions (marked with 'T'). No further preprocessing was applied to the images.

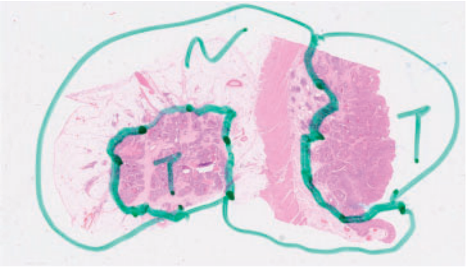


Fig. 1. Typical whole slide image from the data collection. At 10× magnification, this image is $39\,936 \times 22\,528$ pixels in size. The regions marked with a 'T' correspond to tumoral component, while the 'N' annotation indicate normal tissue

1.3 Local descriptors

We based our sample description on the aggregation of local information over the tumor regions in the image. The choice of image features plays a major role in the performance of image recognition/classification system. Traditionally, most of such features are handcrafted, consisting of some dense sampling of local patches, like in wavelet decomposition, Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), Local Binary Patterns (LBP) (Ojala *et al.*, 1996), etc. These local descriptors are later pooled into a global representations by means of methods such as Bag-of-Visual-Words (BoVW) (Csurka *et al.*, 2004), Fisher Vector (FV) (Perronnin and Dance, 2007) or Vector of Locally Aggregated Descriptors (VLAD) (Jégou *et al.*, 2010, 2012).

More recently, Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1989, 2015) gained momentum due to the superior performance of the systems employing them and to the increasing availability of dedicated software (and hardware) systems facilitating their use. While the CNNs also require a number of design decisions (such as their structure), they also have a large number of parameters that are learned from data, leading to adapted image descriptions. Cimpoi *et al.* (2016) provide a detailed comparison of deep image features and some standard ones in the general context of texture classification. In biomedical imaging, there are a number of successful recognition systems based on various CNNs architectures, such as U-Net (Ronneberger *et al.*, 2015). In general, training CNN-based recognition systems requires a large number of labeled image examples, the deeper the architecture more images being needed. For example, the well-known image recognition systems like ImageNet (Krizhevsky *et al.*, 2012) or GoogleNet (Szegedy *et al.*, 2015) were trained on millions of images. Such large data collections are usually not available in biomedical field, thus the interest in transferring general pre-trained CNN models to the medical applications. For example, van Ginneken *et al.* (2015) and Kawahara *et al.* (2016) describe such successful systems that are based on pre-trained CNN features.

An alternate route for obtaining local descriptors is represented by the autoencoding methods, where an identity function is learned under the constraint of a lower dimensional (or sparse) internal representation. The parameters of the function are obtained through an optimization process, where the distance (usually L_2) between the original and reconstructed image is minimized, eventually with some additional constraints over the parameters. Examples of such methods are represented predictive sparse decomposition methods (as used in Chang *et al.*, 2015, for example) and deep autoencoding networks. We do not explore further this direction on the present work.

For the problem addressed here, we chose to use a very deep CNN trained on *ImageNet* data collection—*imagenet-vgg-f* (Chatfield *et al.*, 2014)—as implemented in the MatConvNet library (Vedaldi and Lenc, 2015) (for the architecture see http://www.vlfeat.org/matconvnet/models/imagenet-vgg-f.svg). The network is trained to predict the probability of an input color image of size 224×224 to belong to one of the 1000 categories. By using the output of the next to last layer (*relu7*, before the classification layers), a 4096 element description vector can be obtained. Since we will use Gaussian Mixture Models (GMMs—see Section 1.4) for building the coding dictionary, such a high-dimensional space would require a prohibitively large number of samples for a good fit of the models, so we choose to perform PCA to further reduce the dimension of the local descriptor vectors by retaining the first $d = 128$ coordinates (chosen to be fixed, non-trainable). Thus, a local RGB patch of 224×224 pixels was reduced to a set of 128 values corresponding to the projection of the 4096-value ImageNet vector onto the first 128 principal axes.

As a side note, we remark that the CNN-based descriptor vector is itself the result of a combination of a number of filters applied to even smaller neighborhoods. However, in this work, we consider the basic neighborhood to be the 224×224 patch on which the CNN is applied.

1.4 Aggregating local descriptors

Once a set of local descriptors is obtained from an image, they are pooled into a summarizing feature vector supposed to capture the global aspects of the image. The first step of the process involves the re-coding of the image in terms of elements of a *visual dictionary (codebook)*, the same for all classes, which is followed by the computation of the image representation.

For the construction of the codebook, *k*-means clustering and GMMs are the most common choices, and are typically used with either the standard *Bag-of-Visual-Words* (Csurka *et al.*, 2004) or other aggregators. Jégou *et al.* (2012) give a comprehensive comparison of various design choices. Here we shortly remind the main differences between BoVW, FV and VLAD:

- *Bag-of-Visual-Words* typically uses *k*-means clustering for obtaining a codebook, with the *K* centroids from the clustering being the codewords (visual words). Then the representation of an image is simply the histogram of the number of local descriptors assigned to each codeword, thus an image is reduced to a *K*-dimensional vector. This histogram can be further normalized using Manhattan or Euclidean normalization Jégou *et al.* (2012). One can also use a soft-coding scheme in which the patches are assigned, for example, a code based on the distance to the centroids (Sivic and Zisserman, 2003).
- *Fisher Vector* represents a generalization of BoVW as it encodes higher order statistics of the distribution of the codewords. In this case, the codebook is usually obtained as a GMM with *K* components fitted via expectation maximization on the training data. The FV encodes the gradient of a given sample's likelihood with respect to parameters of the fitted GMM, thus it indicates the direction in the parameter space in which the learned GMM has to be modified to accommodate the observed data (Jégou *et al.*, 2012). For a full FV that accounts for differences both in mean and variance between the model and observed data, the resulting representation vector has $2Kd$ elements (d being the size of the local descriptor vector).
- *VLAD* can be seen as a non-probabilistic version of FV (Jégou *et al.*, 2012) and was designed to provide a low dimensional representation of the image (Jégou *et al.*, 2010) that would allow the indexing of very large image databases in memory. It tries to combine the simplicity of BoVW with some ideas of FV: the codebook is learned via *k*-means clustering and each patch is assigned the closest codeword as in BoVW, but the feature vector accumulates the differences between each patch and its corresponding codeword, similar to FV. See Arandjelovic and Zisserman (2013) for a detailed discussion and further extensions.

In the present work, we decided to use a common method for constructing the visual codebook, namely the GMMs. This allowed us to test a soft-coding scheme as well, in which codes were based on the posterior probabilities of being generated by a particular component of the GMM.

1.5 Classifier training and performance estimation

Training the system could be summarized by the following steps:

1. for each image, extract the local descriptors (based on ImageNet) for all non-overlapping regions corresponding to tumoral component(s);
2. construct a visual codebook by:
   a. performing PCA and retain the first 128 components (the PCA model is saved for later application on validation set)
   b. fitting a $K = 128$-component GMM on PCA-transformed local descriptors (the visual codebook is saved for later usage on validation set)
3. train the binary classifiers (save the models for validation). Each such binary classifier was a support vector machine with a radial basis function kernel. Two parameters were tuned in an inner cross-validation loop: the γ parameter of the kernel and the C parameter for the misclassification penalty. The final prediction of the subtype label is made according to the decision tree in Figure 2. This particular decomposition of the multi-class problem was the result of the analysis of misclassified samples in the development set which suggested that first Subtypes A and B should be separated from the rest (see Section 2.1).

Since the ImageNet is an external model independent of the data analyzed, it does not need to be included in the cross-validation loop, this being an additional reason for preferring a pre-built CNN model. The other steps, however, were repeated at each cross-validation iteration on the corresponding training data.

1.6 Statistical analyses

For the identification of image features enriched/depleted in a subtype with respect to the other subtypes, we used Wilcoxon rank-sum tests since the measurements were not normally distributed. For hierarchical clustering we used the Ward method with an Euclidean distance between feature vectors. Survival analysis was performed using survival package (version 2.39-4) from R statistical computing environment (version 3.3.1, www.r-project.org). The estimation of hazard ratios was obtained from Cox proportional hazards regression in the absence of any other covariates, while the comparison of survival experience of different subgroups was assessed by log-rank test (Mantel–Haenszel test). Statistical significance level was chosen to be $P = 0.01$ and all tests yielding a P value $0.01 \le P \le 0.05$ were considered marginally significant. Finally, the 95% confidence

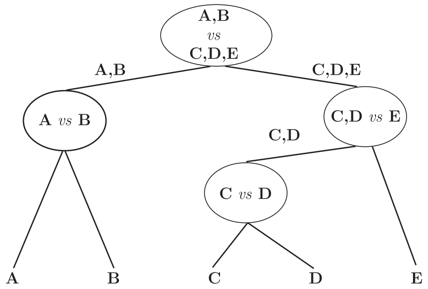


Fig. 2. Decomposition of the multi-class classification problem. For each non-terminal node a binary classifier was trained to split the respective groupings of molecular subtypes

intervals (95%CI) for binomial random variables (such as accuracy) were estimated using the (Agresti and Coull, 1998) method.

2 Results and discussion

The results discussed here are complemented by larger images on the project's website: http://bias.cerit-sc.cz/somopro-subtypes.html.

2.1 Initial experiments

As mentioned, in an attempt to avoid overfitting the available data, a development set has been used to guide the design decisions and to set a number of meta-parameters. We tested dictionaries with $K_1 = 64$ and $K_2 = 128$ codewords and compared the performance of BoVW, FV and VLAD representations when predicting the five molecular subtypes. We performed this comparison under two standard decompositions of the multi-class classification problem, namely *1-vs-all* and *1-vs-1*.

These tests showed that BoVW with GMM-based quantization performed as good as the more involved representation by FV and VLAD see Supplementary Materials—Section S1. The small sample size definitely influences this observation, since both FV and VLAD have much higher dimensionality and would require more data for a better training. Table 1 shows the results for BoVW method with 1-vs-all decomposition of the multi-class problem, on the development set (obtained by stratified 4-fold cross-validation)—for the other approaches the results were similar, so they are not detailed here.

Another important observation was that the 1-vs-1 and 1-vs-all decompositions of the multi-class classification problem might not be the best suited for the present case. By analyzing the confusion matrix and taking into account the performance indexes (precision and recall), it appeared that a first split would have been more advantageous between Classes A and B on one side and C, D and E, on the other side. This observation is also supported by the results in Budinská *et al.* (2013) where it is noted that Subtypes A and B, on one hand, and C, D and E, on the other hand, share dominant and secondary dominant morphological features as well as similar survival expectancy. So, the final design for the multi-class classifier was chosen to be as depicted in Figure 2.

2.2 Prediction of molecular subtypes

Once the final decisions for the classification system were taken based on the initial experiments described above, the performance of the system was assessed using 10-fold cross validation, on the whole set of 300 samples.

The estimated overall accuracy of the multi-class classifier was $\text{Acc} = 0.84$, $95\%\text{CI} = (0.79 - 0.88)$ for a weighted average recall and precision of $R = 0.85$, $95\%\text{CI} = (0.80 - 0.89)$ and $P = 0.84$, $95\%\text{CI} = (0.80 - 0.88)$, respectively. Table 2 details the performance metrics of the classifier. We note the good performance of the first decision level ({A, B} versus {C, D, E}) ($\text{Acc} = 0.89$, $95\%\text{CI} = (0.85 - 0.92)$) but also the poor recognition of the Subtype E.

We repeated the same experiments on the 200 samples not used in the development set and the results were in line with those above (thus not repeated here), only with Subtype A being slightly worse separated from Subtype B (see Supplementary Materials—Section S2). This indicates that the current sample size may still be too small for some cases and some improvements may be expected by enlarging the training set.

2.3 Associations between predictions and clinical data

The study (Budinská *et al.*, 2013) indicated that some associations could be found between molecular subtypes and clinical variables and molecular markers. Hence, we were interested in testing whether such associations are transferrable to the predictions made by the image-based classifier. To avoid overly-optimistic discoveries, we use the predictions (A–E labels) produced during the cross-validation estimation of the system. There is also one caveat: as explained the selection of the cases was governed by technical constraints and thus it does not represent the true population-based statistics for various clinical variables and the results reported here should not be compared directly with those in Budinská *et al.* (2013). Nevertheless, we investigate these associations and compare them with those found between gene expression-based subtypes and the clinical variables, on the same set of cases.

We first tested whether the predicted subtypes were associated with relapse free survival (RFS). In Budinská *et al.* (2013), Subtypes A and B have a lower risk of relapse than Subtypes C, D and E. The same can be observed in the set of 300 samples used here ($P = 0.0014$, $\text{HR} = 1.75, 95\%\text{CI} = (1.24 - 2.49)$, Figure 3(a)). The image-based subtype predictions also produce a statistically significant stratification of the population ($P = 0.012$, $\text{HR} = 1.56, 95\%\text{CI} = (1.10 - 2.21)$, Figure 3(b)).

We also found associations between microsatellite stability, BRAF and KRAS mutations, and mucinous histology and various subtypes—both image-based and gene expression-based. In the case of image-based predictions, Subtypes A and C were enriched in mucinous histology compared with the sample average, while Subtype E was almost depleted of it. BRAF-mutated cases (5.8% of all cases) were mostly found in Subtype C (20% of cases predicted), and rarely in Subtype B (2.4%), while KRAS mutation (38.4% of all cases) represented 77% of cases predicted as Subtype A and only 29% and 22% of cases predicted as Subtypes B and E, respectively. Finally, high microsatellite instability (MSI) was almost exclusively found in Subtype C (10 out of 13 cases). The same trends were found in gene-expression subtypes, with some variations below statistical significance.

Table 1. Confusion matrix for BoVW

	Predicted						
	A	B	C	D	E	Precision	Recall
A	3	4				0.75	0.43
B	1	41		5		0.76	0.87
C		3	7	2		0.44	0.58
D		4	8	13	2	0.59	0.48
E	1	2	1	2	1	0.33	0.14

Empty cells correspond to null values.

Table 2. Ten-fold cross-validation confusion matrix for the multi-class classifier and corresponding per-class performance metrics

	Predicted						
	A	B	C	D	E	Precision	Recall
A	21					0.95	1.00
B	1	119		13	7	0.91	0.85
C		2	29	6		0.91	0.78
D		8	1	71	1	0.75	0.88
E		2	2	5	12	0.60	0.57

Empty cells correspond to null values.

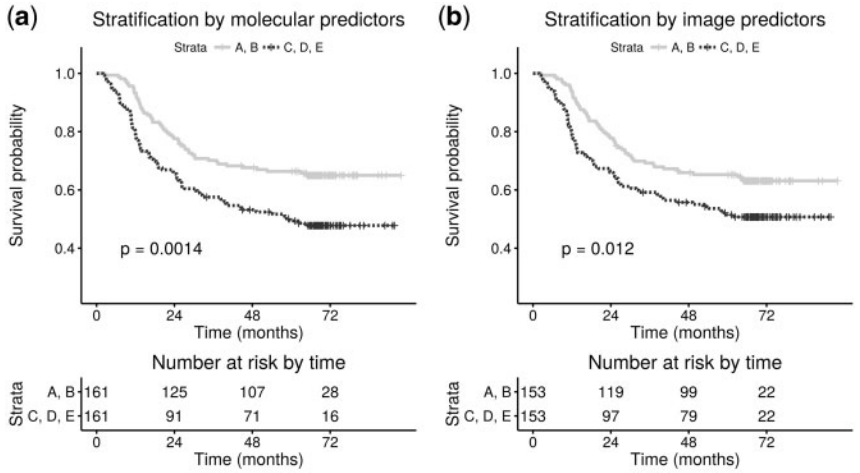


Fig. 3. Survival analysis: risk of relapse stratified by (a) molecular subtypes and (b) image-based classifier. Subtypes A and B represent a lower risk group, while subtypes C, D and E a higher risk

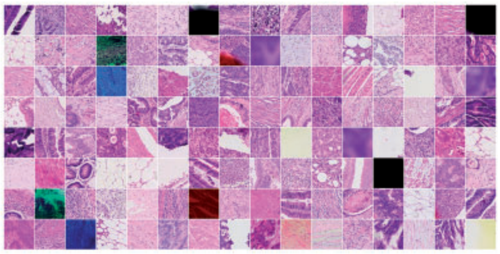


Fig. 4. Visual dictionary for colorectal cancer. While most of the selected visual words correspond to various tissue architectures, some are clearly linked to artifacts still present in the images, or regions partially covered by the annotations. The ordering of the image patches is given by the index in the GMM, with indexes from 0 to 127 (by rows) (see Supplementary Materials - Section S4 for the color version)

A related question was whether the misclassified samples were enriched in any particular type of tumors. The only significant association was between the misclassified Subtype B samples, which were enriched in higher T-stage and N-stage tumors. This observation may provide hints about further refinement of the classifier for Subtype B. Detailed results are given in Supplementary Materials—Section S3.

2.4 Visual codebook

We explored the structure of the visual codebook as obtained by training the model on the full data set. A visual depiction of the extracted codewords (centers of the Gaussian components) is shown in Figure 4 and a higher resolution image is given in Supplementary Materials—Section S4. Note that the visual codewords are the centers of the Gaussians in the GMM, hence the means of feature vectors obtained by projecting the ImageNet features in the PCA space. The patches shown are just the closest image neighborhoods to these centers, thus they are an approximation of the true centers (whose visual appearance would require inverting the CNN function). We use this simplification only for visualization purposes and to get a qualitative assessment of the results.

As one can see most of the codewords could be associated with distinct tissue architectures (from various parts of the glands, papillary or tubular structures, to necrotic and fat regions). On the other hand, it is apparent that some of the codewords were affected—to different degrees—by the markings on the slides. Finally, a few codewords clearly corresponded to artifacts (either due to out-of-focus regions or markings). However, none of these artifact-related codewords were found to be associated with the subtypes, indicating that the approached use can cope, to some extent, with the noise inherent in such images.

Some of the codebooks had a much higher incidence in a particular subtype than in all the others (Wilcoxon rank-sum test). In Figure 5, the top four visual codewords resulted from this analysis are shown along with the corresponding P values (no adjustment for multiple testing was performed, since this is purely exploratory). For all the Subtypes but E, the associations were statistically significant ($P \leq 0.01$). The Subtype E seemed to not have a strong preference for any of the codewords, the few found associations being weakly statistically significant ($0.01 \leq P \leq 0.05$). It appears that Subtype A is associated with well differentiated morphology (Fig. 5(a–d)), with Subtype B being less well differentiated (Fig. 5(e–h)). For Subtypes C, D and E, the top codewords could be associated with either necrotic tissue (Fig. 5 (j and l)), stromal reaction (Fig. 5(m–p)) or poorly differentiated morphology (Fig. 5(q)). It is important to stress that the classifiers were built based on non-linear support vector machines, so the results from this analysis cannot be directly extrapolated to understanding the classification models.

We performed a hierarchical clustering (Ward method) of all the codewords using Euclidean distance and the result showed a rather structured codebook (see Supplementary Materials—Section S5). By corroborating the clustering results with those above, one can see that there are two major clusters—one corresponding mostly to features that are enriched in Subtypes A and B (and depleted in C, D and E) and one corresponding to features enriched in Subtypes C, D and E. This post-hoc analysis supports our decision of having a first decision level separating Subtypes A and B from Subtypes C, D and E.

3 Conclusion

We presented an approach at recognizing the CRC molecular subtypes from the routine histology images. The results indicate that an automated system could be built to identify with high confidence at least four of the five subtypes—Subtype E apparently being much more challenging to recognize. The predictions made by the classifier were found to be also prognostic for relapse-free survival and associated with other clinical parameters, as their molecular counterparts.

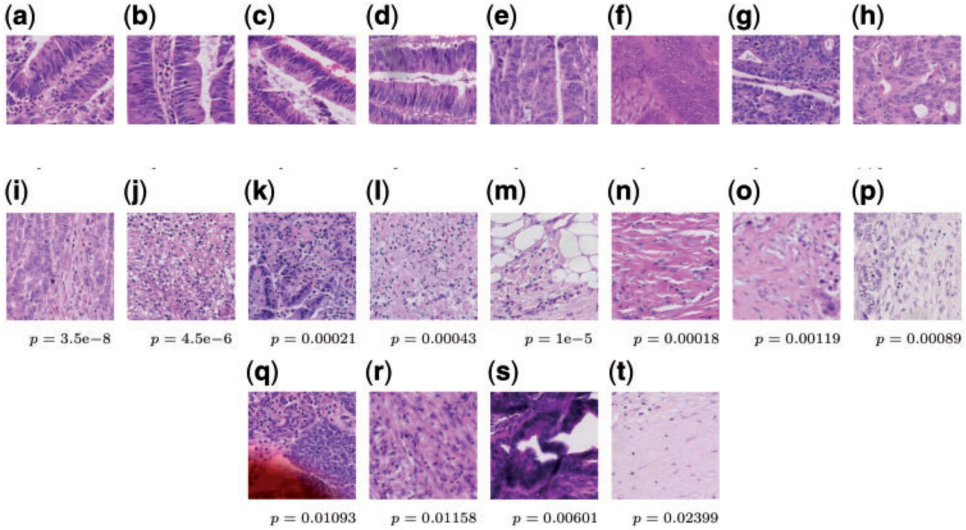


Fig. 5. Top four prototypes associated with each subtype: (a–d) Subtype A, (e–h) Subtype B, (i–l) Subtype C, (m–p) Subtype D and (q–t) Subtype E. Under each image the corresponding *P* value from Wilcoxon rank-sum test is shown

The models used for predicting the subtypes are based on support vector machine classifiers with radial basis functions kernels, making the direct interpretation of the models rather intricate. Nevertheless, we qualitatively evaluated the image features by testing their associations with various subtypes and inspecting their distribution in the whole image. To obtain better insights, we plan to also build simplified models—even at the expense on degraded performance—that would better lend themselves to a biological interpretation, a mandatory condition for the acceptance of the system.

In the current work, we concentrated on recognizing the five molecular subtypes from pre-segmented tumoral regions. This simplification will be addressed in future work where we plan to use an automatic segmentation of the tumor region as a preprocessing step for the subtype recognition. Another question we will address in the future pertains the classification of the so-called ‘outliers’: tumors for which no molecular subtype was assigned. It would be interesting to see how the subtypes predicted by the current image-based classifier correlate with the similarity between their expression profiles and those of well-assigned tumors.

One has to bear in mind that despite recent efforts to consolidate the molecular taxonomy of CRC, the sub-categorization of CRC is still not definitive. Indeed, depending on the size of the cohort and parameters chosen for cut-offs, more or less molecular subtypes can be observed, thus this categorization is still fluid. Nevertheless, in the present work, it has been considered the golden standard to which the image-based models were compared against. We believe that actually combining the observations from the two modalities may lead to an even more refined subtyping of the CRC. However, this would probably involve a more supervised (by expert pathologists) construction of the image-based models.

As they stand now, our results are clearly supporting the possibility of translating some molecular observations into image-based models, as it is the case of molecular subtypes. These results are reinforced by similar observations made by an expert pathologist (Budinská *et al.*, 2013), where several tissue architectural patterns could be linked, in a supervised analysis, to the molecular subtypes. It is interesting to note that some of the regions/patterns found representative in our data-driven analysis are also visually similar to those hand-picked by an expert (see example images in Budinská *et al.*, 2013). On the other hand, the intra-tumoral heterogeneity and pathology sampling region clearly influence sample’s assignment to a molecular subtype (Dunne *et al.*, 2016). In the light of the results presented here, it can be imagined an image-analysis approach to the delineation of the tissue sampling regions to improve the stability of the subtype assignment.

While it is too early for considering any clinical application of the models described here, they could, however, be used for indexing/annotating or for retrieval of samples of interest from archives. Consider the situation in which one would like to test for some biomarker which is hypothesized to work in one or several subtypes on a retrospective collection of samples. Since determining the molecular subtypes relies on profiling hundreds of genes, it makes more sense to use a classifier such the one proposed here, to select the most promising samples. And this can be implemented without significant effort since more and more of the pathology departments are adopting the digital pathology workflows, thus the images being readily available.

Acknowledgements

The necessary computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085 projects under the programme ‘Projects of Projects of Large Research, Development, and Innovations Infrastructures’.

Funding

This project was financially supported from the SoMoPro II programme. The research leading to this result has acquired a financial grant from the People Programme (Marie Curie action) of the Seventh Framework Programme of EU according to the REA Grant agreement no. 291782. The research is further co-financed by the South Moravian Region. This article reflects only the authors’ views and the Union is not liable for any use that may be made of the information contained therein.

Conflict of Interest: none declared.

References

Agner,S.C. *et al.* (2014) Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. *Radiology*, **272**, 91–99.

Agresti,A., and Coull,B.A. (1998) Approximate is better than "Exact" for interval estimation of binomial proportions. *Am. Stat.*, **52**, 119–126.

Arandjelovic,R., and Zisserman,A. (2013). All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, pp. 1578–1585.

Budinská,E. *et al.* (2013) Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.*, **231**, 63–76.

Budinská,E. *et al.* (2016). Experiments in molecular subtype recognition based on histopathology images. In *International Symposium on Biomedical Imaging*. IEEE, Masaryk University, Brno, Czech Republic, pp. 1168–1172.

Chang,H. *et al.* (2011) Morphometic analysis of TCGA glioblastoma multiforme. *BMC Bioinform.*, **12**, 484.

Chang,H. *et al.* (2015) Stacked predictive sparse decomposition for classification of histology sections. *Int. J. Comput. Vis.*, **113**, 3–18.

Chatfield,K. *et al.* (2014). Return of the devil in the details: delving deep into convolutional nets. In *British Machine Vision Conference*.

Cimpoi,M. *et al.* (2016) Deep filter banks for texture recognition, description, and segmentation. *Int. J. Comput. Vis.*, **118**, 65–94.

Cooper,L.A.D. *et al.* (2012) Integrated morphologic analysis for the identification and characterization of disease subtypes. *J. Am. Med. Inf. Assoc.*, **19**, 317–323.

Csurka,G. *et al.* (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 59–74.

De Sousa,E. *et al.* (2013) Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat. Med.*, **19**, 614–618.

Dogan,B.E., and Turnbull,L.W. (2012) Imaging of triple-negative breast cancer. *Ann. Oncol.: Off. J. Eur. Soc. Med. Oncol./ESMO*, **23**, vi23–vi29.

Dunne,P.D. *et al.* (2016) Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clin. Cancer Res.*, **22**, 4095–4104.

Guinney,J. *et al.* (2015) The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, **21**, 1350–1356.

Jégou,H. *et al.* (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, INRIA, Le Chesnay, France, pp. 3304–3311.

Jégou,H. *et al.* (2012) Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, **34**, 1704–1716.

Kawahara,J. *et al.* (2016). Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging*. IEEE, Simon Fraser University, Burnaby, Canada, pp. 1397–1400.

Krizhevsky,A. *et al.* (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114.

Lan,C. *et al.* (2015) Quantitative histology analysis of the ovarian tumour microenvironment. *Sci. Rep.*, **5**, 16317–16317.

LeCun,Y. *et al.* (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, **1**, 541–551.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Li,G. *et al.* (2016) Embracing an integromic approach to tissue biomarker research in cancer: Perspectives and lessons learned. *Brief. Bioinform.*, doi: 10.1093/bib/bbw044.

Lowe,D.G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*. The University of British Columbia, Vancouver, Canada, pp. 1150–1157.

Marisa,L. *et al.* (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.*, **10**, e1001453.

Ojala,T. *et al.* (1996) A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.*, **29**, 51–59.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Perronnin,F., and Dance,C. (2007). Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Xerox Research Centre Europe, Meulan, France, pp. 1–8.

Roepman,P. *et al.* (2013) Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int. J. Cancer*, **134**, 552–562.

Ronneberger,O. *et al.* (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer International Publishing, Cham, pp. 234–241.

Sadanandam,A. *et al.* (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.*, **19**, 619–625.

Satyanarayanan,M. *et al.* (2013) OpenSlide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.*, **4**, 27.

Sivic,J., and Zisserman,A. (2003). Video google: a text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. University of Oxford, Oxford, United Kingdom, pp. 1470–1477.

Stålhammar,G. *et al.* (2016) Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathol.*, **29**, 318–329.

Szegedy,C. *et al.* (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, pp. 1–9.

Van Cutsem,E. *et al.* (2009) Randomized phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J. Clin. Oncol.*, **27**, 3117–3125.

van Ginneken,B. *et al.* (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI*. IEEE, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands, pp. 286–289.

Vedaldi,A., and Lenc,K. (2015). MatConvNet – convolutional neural networks for MATLAB. In *ACM International Conference on Multimedia*, pp. 1–55.

Weigelt,B. *et al.* (2010) Histological types of breast cancer: how special are they? *Molecular Oncology*, **4**, 192–208.

[*14*] **Budinská E**, Hrivňáková M, Ivkovic TC, Madrzyk M, Nenutil R, Bencsiková B, Al Tukmachi D, Ručková M, Zdražilová Dubská L, Slabý O, Feit J, Dragomir MP, Borilova Linhartova P, Tejpar S, Popovici V. Molecular portraits of colorectal cancer morphological regions. Elife. 2023 Nov 13;12:RP86655. doi: 10.7554/eLife.86655. PMID: 37956043; PMCID: PMC10642970.

RESEARCH ARTICLE

Molecular portraits of colorectal cancer morphological regions

Eva Budinská[1], Martina Hrivňáková[1], Tina Catela Ivkovic[2], Marie Madrzyk[2], Rudolf Nenutil[3], Beatrix Bencsiková[3], Dagmar Al Tukmachi[2], Michaela Ručková[2], Lenka Zdražilová Dubská[4], Ondřej Slabý[5], Josef Feit[6], Mihnea-Paul Dragomir[7,8,9], Petra Borilova Linhartova[1], Sabine Tejpar[10], Vlad Popovici[1]*

[1]RECETOX, Faculty of Science, Masarykova Univerzita, Brno, Czech Republic; [2]Central European Institute of Technology, Masarykova Univerzita, Brno, Czech Republic; [3]Masaryk Memorial Cancer Institute, Brno, Czech Republic; [4]Faculty of Medicine, Masarykova Univerzita, Brno, Czech Republic; [5]Central European Institute of Technology, Department of Biology, Faculty of Medicine, Masarykova Univerzita, Brno, Czech Republic; [6]Department of Pharmacology and Toxicology, Faculty of Pharmacy, Masarykova Univerzita, Brno, Czech Republic; [7]Institute of Pathology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany; [8]Berlin Institute of Health, Berlin, Germany; [9]German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany; [10]Faculty of Medicine, Digestive Oncology Unit, Katholieke Universiteit Leuven, Leuven, Belgium

***For correspondence:** vlad.popovici@recetox.muni.cz

Competing interest: The authors declare that no competing interests exist.

Funding: See page 16

Preprint posted
01 February 2023
Sent for Review
23 February 2023
Reviewed preprint posted
28 June 2023
Reviewed preprint revised
24 October 2023
Version of Record published
13 November 2023

Reviewing Editor: Yongliang Yang, Dalian University of Technology, China

Abstract Heterogeneity of colorectal carcinoma (CRC) represents a major hurdle towards personalized medicine. Efforts based on whole tumor profiling demonstrated that the CRC molecular subtypes were associated with specific tumor morphological patterns representing tumor subregions. We hypothesize that whole-tumor molecular descriptors depend on the morphological heterogeneity with significant impact on current molecular predictors. We investigated intra-tumor heterogeneity by morphology-guided transcriptomics to better understand the links between gene expression and tumor morphology represented by six morphological patterns (morphotypes): complex tubular, desmoplastic, mucinous, papillary, serrated, and solid/trabecular. Whole-transcriptome profiling by microarrays of 202 tumor regions (morphotypes, tumor-adjacent normal tissue, supportive stroma, and matched whole tumors) from 111 stage II-IV CRCs identified morphotype-specific gene expression profiles and molecular programs and differences in their cellular buildup. The proportion of cell types (fibroblasts, epithelial and immune cells) and differentiation of epithelial cells were the main drivers of the observed disparities with activation of EMT and TNF-α signaling in contrast to MYC and E2F targets signaling, defining major gradients of changes at molecular level. Several gene expression-based (including single-cell) classifiers, prognostic and predictive signatures were examined to study their behavior across morphotypes. Most exhibited important morphotype-dependent variability within same tumor sections, with regional predictions often contradicting the whole-tumor classification. The results show that morphotype-based tumor sampling allows the detection of molecular features that would otherwise be distilled in whole tumor profile, while maintaining histopathology context for their interpretation. This represents a practical approach at improving the reproducibility of expression profiling and, by consequence, of gene-based classifiers.

eLife assessment

This study presents a **valuable** finding on the putative molecular patterns underlying characteristic morphological regions observed in colorectal cancer (CRC). The authors provide a morphological framework through which clinicians might improve the performance of molecular signatures and consequently predict the clinical response of patients with better accuracy. The evidence supporting the claims of the authors is **solid**. The work will be of interest to clinicians and cancer biologists working in the field of CRC.

Introduction

Colorectal cancer (CRC), the third cause of death among cancer patients, is a highly heterogeneous disease, with a slow initial progression that favors the accumulation of mutations leading to a complex phenotype. Differences that exist both between and within tumors of the same cancer type are a major hurdle towards proper treatment selection and for developing more targeted therapies. Depending on the perspective under which these differences are investigated, various categorization paradigms have emerged. The systematization of clinical and histopathological parameters led to the definition of current TNM staging system (***Amin, 2017***), which presently constitutes the gold standard for diagnosis and prognosis. The development of high throughput molecular technologies brought a novel perspective and set the stage for the appearance of molecular taxonomies categorizing the tumors into subgroups sharing common molecular traits (***Perez-Villamil et al., 2012***; ***Budinska et al., 2013***; ***Marisa et al., 2013***; ***De Sousa E Melo et al., 2013***; ***Sadanandam et al., 2013***; ***Roepman et al., 2014***) with consensus molecular subtypes (CMS)(***Guinney et al., 2015***) representing their common denominator. While these studies were based on whole-tumor (bulk) gene expression data, the developments in single-cell sequencing further refined the CMS classes adding two intrinsic epithelial subtypes (iCMS2/3) to the picture (***Joanito et al., 2022***). Other studies combined genomics and transcriptomics data and an alternative classification emerged (***Muzny et al., 2012***).

Whole transcriptome expression profiling of tissue sections is generally performed on RNA extracted from regions of interest covering diverse cell collections. By consequence, the expression levels associated with various transcripts represent, in the end, a weighted mean of contributions of each cell type, being driven by the most abundant ones. The signals from less abundant cell types are reduced or even silenced and are, therefore, overlooked. In the case of solid tumors, this approach requires a representative region, enriched in tumoral cells, to be selected in the tissue section(s) and used for RNA extraction. This is the predominant approach to tissue expression profiling that fueled the myriad of studies over the last two decades and led to significant progress in understanding the various cancers. Newer technologies such as single cell sequencing and spatial transcriptomics allow for a much finer selection of cells to be interrogated (***Tang et al., 2019***; ***Rao et al., 2021***). However, while powerful, these techniques rely on fresh tissue and have still to find their place in routine clinical practice.

The importance of a morphological perspective on the molecular classification has been acknowledged from the beginning, ***Jass, 2007*** already identifying several morphological features associated with the five groups proposed (e.g. serration, mucinous and poor differentiation were highly present in two of the five groups), but also noted that these features were not sufficient for predicting the groups. Later, ***Budinska et al., 2013*** proposed six morphological patterns (morphotypes) as major histological descriptors and showed that a two-tier histological score is strongly associated with the five molecular subtypes identified. Interestingly, a pure data-driven image-based classifier for the same molecular subtypes resulted in selecting remarkably similar morphological motifs (***Budinska et al., 2016***; ***Popovici et al., 2017***). ***Müller et al., 2016*** reviewed the TCGA and CMS subtypes and their links with some morphological aspects, most notably the serrated phenotype. It is worth mentioning that in all these cases the evaluation of the morphological features referred to the whole tumor section; for example, a tumor was considered of mucinous morphology if the mucinous pattern was present in more the 50% of the tumor region, in accordance with standard definitions endorsed by the World Health Organization (***Bosman, 2010***). These links between tumor morphology and molecular features also imply that the gene expression profile may depend on the tumor region sampled for RNA extraction. The sensitivity of gene-based classifiers to tumor sampling raised concerns regarding the stability of consensus molecular subtypes (***Dunne et al., 2016***) and may partially explain the low proportion of biomarkers that reach clinical relevance (***Stewart et al., 2017***).

It is evident that, while intra-tumoral heterogeneity is recognized as a major challenge, we still lack the practical tools for its characterization that would easily translate into a diagnostic and predictive model. In contrast with previous results, in our study we explored region- (morphotype-) based transcriptomics approach as a possible solution to this problem. This method offers a trade-off between whole-tumor profiling and spatial transcriptomics. It has a better signal resolution than whole-tumor profiling, since it selects tumor regions with more similar cellular buildup, and covers the whole transcriptome, but clearly has a much lower spatial resolution than true spatial transcriptomics. However, it represents a practical approach where several regions of interest can be stably identified, and their profiling could be easily integrated in the current molecular pathology diagnostic practice.

Building on our previous results (***Budinska et al., 2013***), we based our study on a detailed exploration of the transcriptome of the six morphotypes identified earlier as associated to the molecular subtypes of CRC: complex tubular (CT), desmoplastic (DE), mucinous (MU), papillary (PP), serrated (SE) and solid/trabecular (TB), respectively. As reference, we also profiled several tumor-adjacent normal (NR) and supportive stroma (ST) regions. The present study was based on a single center cohort and was designed to achieve several goals: (i) identifying representative samples for each of the morphotypes, (ii) providing a comprehensive characterization of their transcriptomics landscape, and (iii) studying the intra-tumoral heterogeneity from the perspective of morphotype-resolved transcriptomics. We characterized the morphotypes from several transcriptional angles: basic molecular programs as captured by differential expression and pathway analyses, molecular tumor classifiers and prognostic gene signatures. At the same time, we looked for variations both across all tumors and across matched samples (within tumors). The emerging picture is of an unexpectedly high heterogeneity, with clear implications both from fundamental biological and practical perspectives, opening new avenues for biomarker design.

Results

Data

From n=111 unique cases of primary CRC tumors (stages: II: 59, III:32, IV:20), n=202 regions were macrodissected representing either tumor morphological regions (n=149), tumor-adjacent normal tissue (NR, n=17), supportive stroma (ST, n=8), or whole tumor (n=28), respectively. Among the tumor morphological regions, n=126 'core samples' were identified based on 'morphological purity', indicating regions containing at least 80% of a unique morphological pattern. The six morphotypes of interest (***Figure 1***) consisted of (in brackets the additional non-core samples) 41 (+11) CT, 13 (+2) DE, 18 (+3) MU, 10 (+2) PP, 33 (+7) SE, and 9 TB samples, respectively. The distribution of associated main clinical parameters is given in ***Supplementary file 1***. The only statistically significant associations found were between MU or TB and grade 3 tumors, and SE and lower grade tumors (p=0.019, ***Supplementary file 2***), respectively.

To complement the results presented here, we created a web application https://morphogene.recetox.cz allowing the interrogation of gene expression in various morphological regions.

Morphotype cellular admixtures

The transcriptomic profile of solid tumor sample is a mixture of gene expression profiles of individual cell types and their specific programs, including cancer cells at different levels of differentiation, specific immune cells, or supportive fibroblasts. As a first step, we performed in-silico deconvolution of the expression profiles to identify the most prevalent cell types in each of the morphotypes and GSEA to score cell-type-specific gene sets (see Materials and methods) in each morphotype, and NR and ST regions (used as controls, ***Figure 2***, ***Supplementary files 3–4***).

The results from ESTIMATE indicated, as expected, a high stromal content for ST, DE, and MU and a high epithelial tumor cell content for normal region and TB and SE morphotypes, respectively (***Figure 2A***). A more balanced situation was observed for CT and PP morphotypes (similar to NR). This agreed with the (stroma-related) 'Isella signatures' (***Isella et al., 2015***) where ST, DE, and MU were enriched in endothelial cells, CAFs and immune cells (***Figure 2B***). When investigating the categories of epithelial cells, the signatures of top of the normal colon crypt cells (***Kosinski et al., 2007***) and

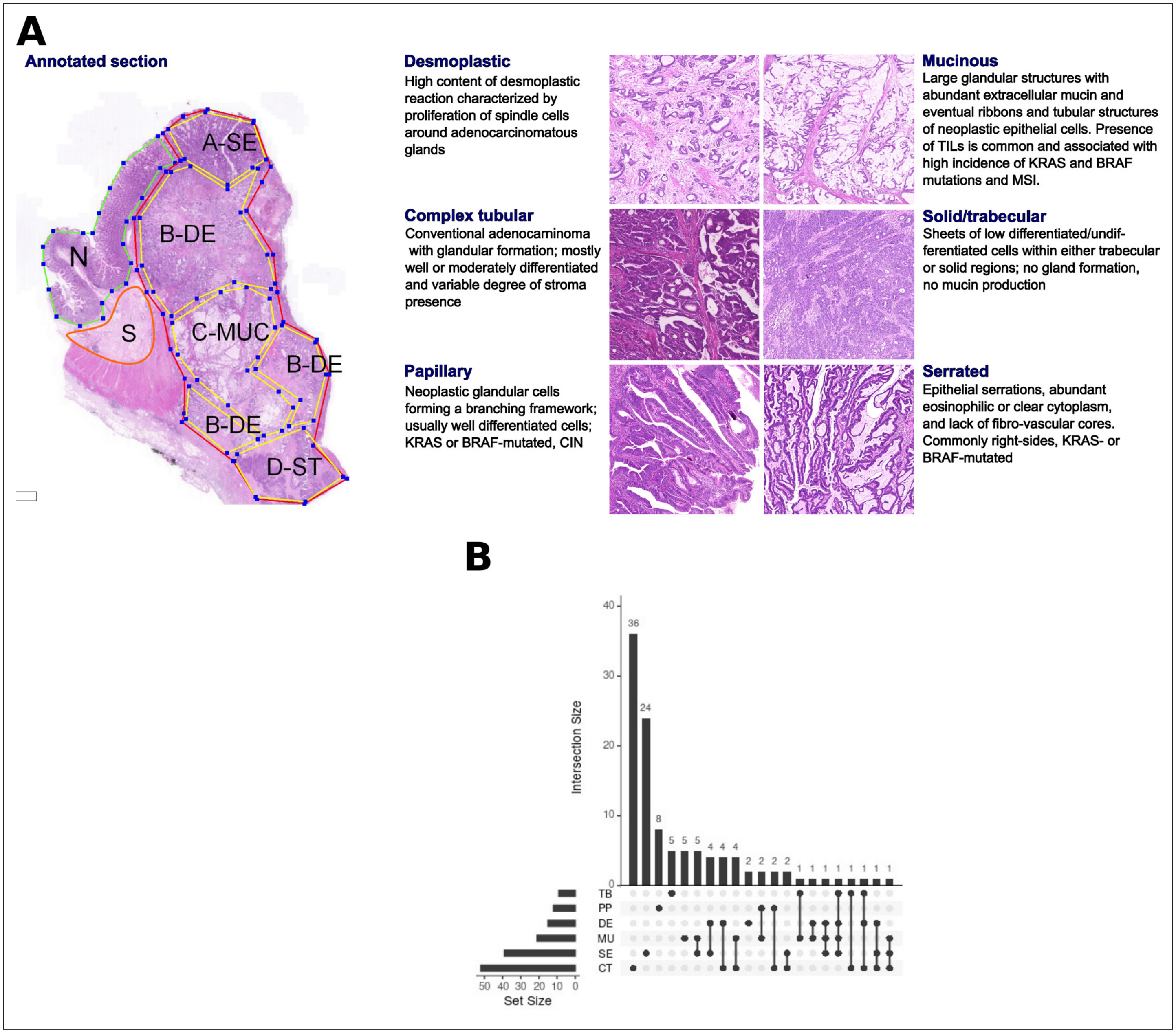

A
Annotated section
A-SE
B-DE
N
S
C-MUC
B-DE
B-DE
D-ST
Desmoplastic
High content of desmoplastic reaction characterized by proliferation of spindle cells around adenocarcinomatous glands
Mucinous
Large glandular structures with abundant extracellular mucin and eventual ribbons and tubular structures of neoplastic epithelial cells. Presence of TILs is common and associated with high incidence of KRAS and BRAF mutations and MSI.
Complex tubular
Conventional adenocarninoma with glandular formation; mostly well or moderately differentiated and variable degree of stroma presence
Solid/trabecular
Sheets of low differentiated/undif-ferentiated cells within either trabecular or solid regions; no gland formation, no mucin production
Papillary
Neoplastic glandular cells forming a branching framework; usually well differentiated cells; KRAS or BRAF-mutated, CIN
Serrated
Epithelial serrations, abundant eosinophilic or clear cytoplasm, and lack of fibro-vascular cores. Commonly right-sides, KRAS- or BRAF-mutated
B
Intersection Size
40
30
20
10
0
36
24
8
5 5 5
4 4 4
2 2 2 2
1 1 1 1 1 1 1 1
TB
PP
DE
MU
SE
CT
50 40 30 20 10 0
Set Size

Figure 1. Morphological patterns and their distribution in the dataset. (**A**) The six CRC morphological patterns of interest (morphotypes). *Left*: example of an original annotation used for macrodissection and RNA extraction. Note that the original annotations in the image are not identical to the ones used in the main text. Here, A-SE stands for serrated (SE) in the text, B-DE for desmoplastic (DE) in the text, C-MUC for mucinous (MU) in the text, and D-ST for solid/trabecular (TB) in the text, respectively. Also, N indicates a tumor-adjacent normal epithelial region and S a supportive stroma region, respectively. *Right*: examples of morphotypes – complex tubular (CT), desmoplastic (DE), mucinous (MU), papillary (PP), serrated (SE), and solid/trabecular (TB). (**B**) Morphotype distribution per case (unique tumor) and intersections thereof: some cases had several morphotypes profiled.

colon differentiated epithelial cells (***Merlos-Suárez et al., 2011***) were enriched solely in NR regions, while DE, MU, CT, SE, and TB were depleted in these cell types (***Figure 2B***). On the other hand, MU, CT, PP, and TB regions expressed genes specific for the basal crypt cells (***Kosinski et al., 2007***) and ST, DE, and MU were enriched in signatures of intestinal stem cells. These observations are in perfect agreement with the definition of the morphotypes and confirm the proper selection of the samples. quanTIseq revealed that all tumor morphotypes were enriched in M1 macrophages (with maximal presence in MU and DE), while M2 macrophages, NK cells and myeloid dendritic cells where

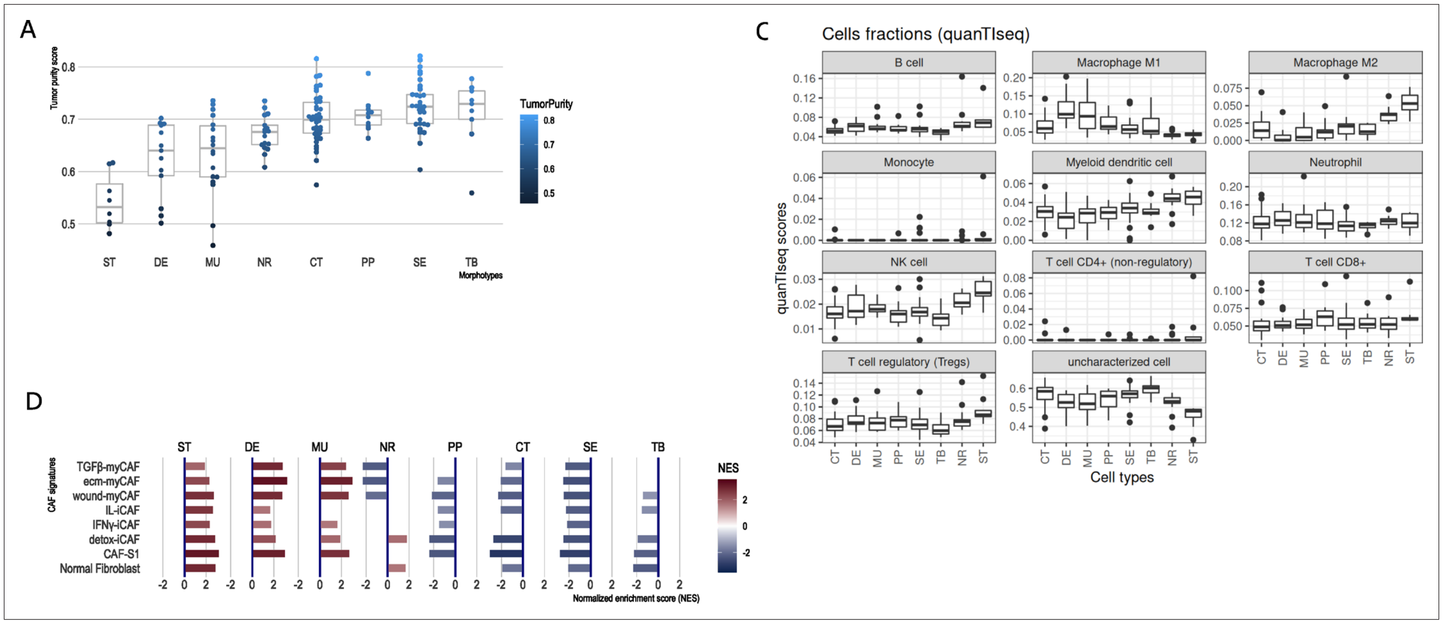


Figure 2. CRC morphotypes: in silico decomposition of the cellular admixture. (**A**) Boxplots of the tumor purity (epithelial content – ESTIMATE method) in each tumor morphotype and the two non-tumor regions, ordered by increasing median values. (**B**) Signatures specific to colon crypt compartments and major cell types estimated from gene expression data in terms of normalized enrichment scores (NES): only statistically significant scores are shown. (**C**) Immune cell fractions (and unassigned fractions) inferred from gene expression data using quanTIseq method. (**D**) Types of cancer-associated fibroblasts (CAFs) as estimated from gene expression using the signatures from ***Khaliq et al., 2022***; ***Kieffer et al., 2020***.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Epithelial signatures from ***Pelka et al., 2021***.

Figure supplement 2. Immune signatures from ***Pelka et al., 2021***.

Figure supplement 3. Stromal signatures from ***Pelka et al., 2021***.

highly present in supporting stroma and tumor-adjacent normal regions (***Figure 2C***). Additionally, TB morphotype had the lowest scores for regulatory T cells (TREGs) and B cells.

Further we refined the morphotype cell admixtures by testing signatures of different cell types and their active programs as derived from single-cell sequencing studies. We evaluated more than 150 signatures of stromal, epithelial, and immune cell population (supplemental tables of ***Pelka et al., 2021***) and cancer associated fibroblasts (CAFs) (***Khaliq et al., 2022***; ***Kieffer et al., 2020***) (see ***Supplementary files 3–4*** for full signatures). Interestingly, the morphotypes differed in the signatures of CAFs subpopulations (***Figure 2D***). ST, MU, and DE had high GSEA scores of most of the CAFs subpopulations, while the rest (CT, PP, SE, and TB) had mostly negative scores, indicative of depletion of corresponding cell types. DE and MU were most strongly enriched in signatures of ECM-myCAF S1 – associated with immunosuppressive microenvironment and pro-metastatic functions (***Kieffer et al., 2020***) – and wound-healing myCAF S1 populations, while the adjacent stroma mainly showed signatures of normal fibroblasts, detox-iCAF S1 and IL-iCAF S1 populations, both characterized by detoxification and inflammatory signaling. NR regions were enriched only in normal fibroblasts and detox-iCAF S1. By exploring the signatures from ***Pelka et al., 2021***, we observed even finer differences between morphotypes within all three cell type populations and their programs (***Figure 2—figure supplements 1–3***). For instance, CT, TB, and SE had enriched pS04 (ribosomal) and pS12 (proliferation) stromal cell signatures, in addition, CT and TB expressed pS05 (interferon-stimulated genes, ISGs) and pS21 (*FOS*, *JUN*) signatures. Also, NR had a specific enrichment in mitochondrial (pS09), metallothionein (pS16) and BMP-producing (pS17) fibroblasts. CT and TB resembled MU in expressing pS20 signature and, additionally, TB showed similar levels of pS13 (inflammatory) signature as MU and DE. ST regions and DE and MU morphotypes had significantly increased pS02 (Fibro. matrix/stem cell niche) signature. Full results for other cell types and programs are provided in the ***Supplementary file 4***.

Table 1. Results of comparison of each morphotype (and the two non-tumoral regions) with the average profile.

The table shows the top 20 up- and down- regulated genes and significantly activated hallmark pathways and processes (as result of GSEA). The genes not significant after p-value adjustment (at FDR = 0.15) have their symbols greyed. See also ***Supplementary files 5–6***.

Morph	Top 20 up-regulated genes (compared to mean)	Top 20 down-regulated genes (compared to mean)	Hallmark pathways with high score	Active processes (based on the active hallmark pathways)
MU	*ARF4, MUC2, SULF1, FNDC1, LOXL1, LGALS1, ANTXR1, BGN, COL12A1, PALLD, MEG8, DKK3, ACVR1, GPX8, CALD1, FBN1, MLLT11, CSRP2, TUSC3, GREM1*	*TIMD4, PRELID3BP3, EREG, KDM4A, CCDC175, TDP2, CHMP1B2P, ACE2, NLRP7, UGT2A3, SLC26A3, A1CF, TSPAN6, CLDN10, TMIGD1, BMP5, MS4A12, FAM3B, CLCA4, MEP1A*	EMT, TNF a signaling via NFKB, Complement, IL2 STAT5 signaling, hypoxia, inflammatory response, KRAS signaling, UV response, myogenesis, coagulation, apical junction, allograft rejection, IL6 JAK STAT3 signaling, interferon gamma response, apoptosis, TGF-beta signaling, angiogenesis, hedgehog signaling, estrogen response early, NOTCH signaling, WNT beta catenin signaling, cholesterol homeostasis	Inflammation, neoangiogenesis, increased metastatic potential, apoptosis, development
DE	*OLFML2B, INHBA, LUM, SULF1, PTPN14, PRDM6, SPOCK1, RDX, EDNRA, COL12A1, CTHRC1, PRRX1, LGALS1, COPZ2, COL10A1, TNFAIP6, IGFL1P1, ST6GAL2, FAP, BGN*	*SLC17A4, ANPEP, DEFA5, RAP1GAP, MRAP2, ADH1C, TRIQK, REG1A, SLC4A4, UGT2B15, REG4, SEMA6A, L1TD1, MS4A12, SI, SPINK4, CLCA4, MUC2, CLCA1, CA1*	EMT, TNF a signaling via NFKB, Complement, IL2 STAT5 signaling, hypoxia, inflammatory response, KRAS signaling, UV response, myogenesis, coagulation, apical junction, apoptosis, TGF-beta signaling, angiogenesis, hedgehog signaling, estrogen response early	Inflammation, neoangiogenesis, increased metastatic potential, apoptosis
PP	*PTPRD, KNDC1, MIMT1, UPK3B, MPZ, MMP15, CYP4F12, SNORD4A, SNAR-C3, TMTC4, LRCOL1, GATA5, SNAR-E, EPHA7, IPO4, SNAR-I, CASC21, NUTF2, SNAR-B2, RPL31P50*	*IGKV3-11, IGHV4-39, ANPEP, OR4F8P, HEPACAM2, ADAM28, CPS1, TMIGD1, NPY6R, ITLN1, SI, ADH1C, CAV1, MMP2, FDCSP, CLU, REG1A, RSPO3, PAX8-AS1, PALMD*	MYC targets V1, MYC targets V2, E2F targets, KRAS signaling DOWN, WNT beta catenin signaling,	
SE	*PPAN-P2RY11, TUBB4BP7, JADE3, PFDN6, CLDN2, YAF2, BOLL, SLAMF9, SLC12A2, CCDC175, GRIN2B, TUBB3P2, GAPDHP71, RPS2P25, MAT1A, NOX1, SNORD12C, SMAD6, MECOM, EXTL2*	*IGKV2D-29, MYLK, TAGLN, CNTNAP3P2, GLI3, CPXM2, NR3C1, CNN1, PECAM1, COLEC12, IGKV4-1, IGKV2D-30, DPYD, CLU, TSHZ2, ADH1B, IL10RA, PDE7B, ABCA8, CDC42SE2*	MYC targets V1, MYC targets V2, E2F targets, G2M checkpoint,	
CT	*TMEM97, RPL13, CLDN1, TFDP1, CKS2, CDCA7, TPX2, ANLN, RAD54B, KRT18, HSPH1, CCT6A, PLK1, TMEM97P2, CSE1L, MIPEP, SNORA71D, SNORA71C, PTTG1, PLBD1*	*CR2, OGN, SNORD114-21, SLC30A10, CLCA4, SNORD114-12, DCLK1, FAT4, CPA3, ADH1B, SLC26A2, SNORD114-20, SFRP1, ZG16, FGF7, SNORD113-1, ABCA8, B4GALNT2, MS4A12, CA1*	MYC targets V1, MYC targets V2, E2F targets, G2M checkpoint, MTORC1 signaling, unfolded protein response, Glycolysis, oxidative phosphorylation, fatty acid metabolism, protein secretion	Proliferation, Catabolism, oxidative stress, cell cycle disruption
TB	*CKAP2, HSP90AA1, PPP3CA, REEP4, MSH6, TOP2A, HSPE1, PPP2R5C, TBCA, VRK2, NIFK, TXNL4A, MNAT1, ERI1, XPO1, VTRNA1-2, ANP32A, ARF6, RNF2, EIF4A1P7*	*FLJ22763, TMEM236, NPY6R, IGKV3D-20, IGKV2D-30, OLFM4, SELENBP1, LRRC19, CDHR1, IGHA1, SNORD123, SLC26A3, CXCL14, SLC3A1, SEMA5A, MS4A12, IGHA2, CLCA4, NXPE4, NXPE1*	MYC targets V1, MYC targets V2, E2F targets, G2M checkpoint, MTORC1 signaling, unfolded protein response, Glycolysis, oxidative phosphorylation, fatty acid metabolism, protein secretion, cholesterol homeostasis,	Inflammation, catabolism, apoptosis, oxidative stress, proliferation, cell cycle disruption
NR	*PIGR, SLC26A3, ADH1B, NXPE1, IGHA2, CLCA1, JCHAIN, IGHA1, FCGBP, IGK, NXPE4, SLC9A2, MUC2, NR3C2, TMEM236, MS4A12, FABP1, IGLC3, IGKV1D-39, LRRC19*	*TACSTD2, FAM83D, ASPN, CXCL11, CTHRC1, SLC39A6, IFNE, SULF1, HSPH1, ELFN1-AS1, THBS2, CLDN1, SIM2, SLC22A3, SPARC, FN1, AHNAK2, COL11A1, SPP1, INHBA*	Heme metabolism, bile acid metabolism, xenobiotic metabolism, fatty acid metabolism	
ST	*SFRP2, ADH1B, EMCN, STEAP4, ADAMTS1, ABI3BP, SPARCL1, DCN, PTGDS, PALMD, NOVA1, SLIT3, OGN, SERPINF1, RSPO3, CPA3, FBLN5, C3, EFEMP1, PBX3*	*FRK, AADACP1, CKS2, HOOK1, CLDN1, ANLN, S100P, UGT8, MACC1, EXPH5, CYP3A5, OCIAD2, SLC12A2, GK, EVADR, TMC5, REG4, TFF1, TCN1, CXCL8*	EMT, TNF a signaling via NFKB, Complement, IL2 STAT5 signaling, hypoxia, inflammatory response, KRAS signaling, UV response, myogenesis, coagulation, apical junction, allograft rejection, IL6 JAK STAT3 signaling, interferon gamma response	Inflammation, neoangiogenesis, increased metastatic potential

CRC morphotypes and molecular programs

The molecular programs and pathways represented in MSigDB were scored by performing GSEA on differentially expressed genes (DEGs) in all morphotypes (and NR and ST).

For the first analysis, the ordered lists of DEGs per morphotype were obtained by contrasting the individual expression profiles to the average profile of pooled samples (***Supplementary file 5*** contains all DEGs). This allowed the identification of all molecular programs significantly de-/activated in each morphotype (***Table 1***, ***Figure 3***; ***Supplementary file 6***). When considering only the hallmark signatures (H collection), the discriminative gradients between the morphotypes (and NR and ST) were along the EMT and TNF-α signaling axes at one end, and the *MYC* and E2F targets at the other end (***Table 1***, ***Figure 3A***). Desmoplastic and mucinous shared active pathways involved in immune system response (TNF-α signaling via NF-κB, interferon gamma response, complement, *IL2-STAT5* signaling), neoangiogenesis, and increased metastatic potential (EMT, coagulation, TGF-β, NF-kB, NOTCH, Apical junction). At the other end of the spectrum, CT and TB morphotypes had activated major pathways involved in proliferation processes (P53, MTORC 1, Myc targets, G2M checkpoint, Mitotic spindle, NOTCH signaling, Protein secretion). In contrast with CT, TB morphotype shared with MU and DE active TGF-β signaling, apoptosis, and most pathways involved in immune system response. PP and SE morphotypes had activated *MYC* and E2F targets, with PP morphotype exhibiting downregulation of the *KRAS* signaling and upregulation of the WNT-β catenin signaling.

We performed principal component analysis (PCA) of the GSEA scores of hallmark pathways. Their projection onto the first two principal components revealed a specific bi-dimensional clustering of the morphotypes and illustrated the gradient of changes between morphotypes (***Figure 3B***, ***Figure 3—figure supplement 1***). At one end, MU and DE shared the same region in PCA space with positive coordinates on the axis defined, among others, by EMT, inflammatory response, and UV response. At the same time, they had opposite projections on the second axis of variation, defined by p53, unfolded protein response and cholesterol homeostasis. In contrast, SE and PP shared the same quadrant with negative coordinates on the first axis, but positive on the second axis. The CT and TB fell between the two previous groups with respect to the first axis of variation, while having similar activations of pathways defining the principal components. Overlaid on top of the transcriptomics layer, an additional gradient could be observed: epithelial cell differentiation. Indeed, while SE, PP, and CT were well or moderately differentiated, TB, DE and MU had low or undifferentiated morphology.

Figure 3C shows heatmap of median expressions of all top 5 up- and down-regulated genes of each morphotype with respect to the average profile (full lists in ***Supplementary files 5–6***). A second analysis identified morphotype-specific processes and pathways by GSEA of differentially expressed genes between each morphotype and all other five (excluding ST and NR) (***Supplementary files 7–8***).

Several macrodissected regions originated from the same section allowing for paired comparison of morphotypes. While the reduced number of such pairs (MU vs SE: 8 pairs, DE vs SE: 7, CT vs DE: 5, and CT vs MU: 5, respectively) impacted the statistical power, we were able to identify genes differentially expressed (after p-value adjustment) in all but MU vs SE, indicative of regional differences (***Supplementary files 9–10***). The differences between gene expression signatures from the matched paired comparisons were in line with those from comparisons not accounting for sample pairing, indicating that the morphotype specific effect was dominating the contrasts (see ***Figure 3—figure supplement 2***).

We also performed comparison between all pairs of morphotypes (***Supplementary files 11–12***). This comparison shows that, despite similar content in terms of fibroblasts or epithelial cells (discussed above), there are still differences both in terms of differentially expressed genes (***Supplementary file 11***) and activated molecular programs (***Supplementary file 12***) between DE and MU, on one side, and CT, PP, SE, and TB. These results refine those presented above and allow an ordering of morphotypes in terms of relative activation of pathways. For example, *KRAS* signaling appears to be highest in PP, followed by CT.

Morphotypes and molecular subtypes

The molecular subtyping taxonomies of CRC were derived from datasets representing profiles of whole tumor sections, therefore aggregating the expression of many cell types. In our previous work (***Budinska et al., 2013***), we associated molecular subtypes with morphotypes assessed on the whole tumor and hence we were interested to see how this observation translated to the case of

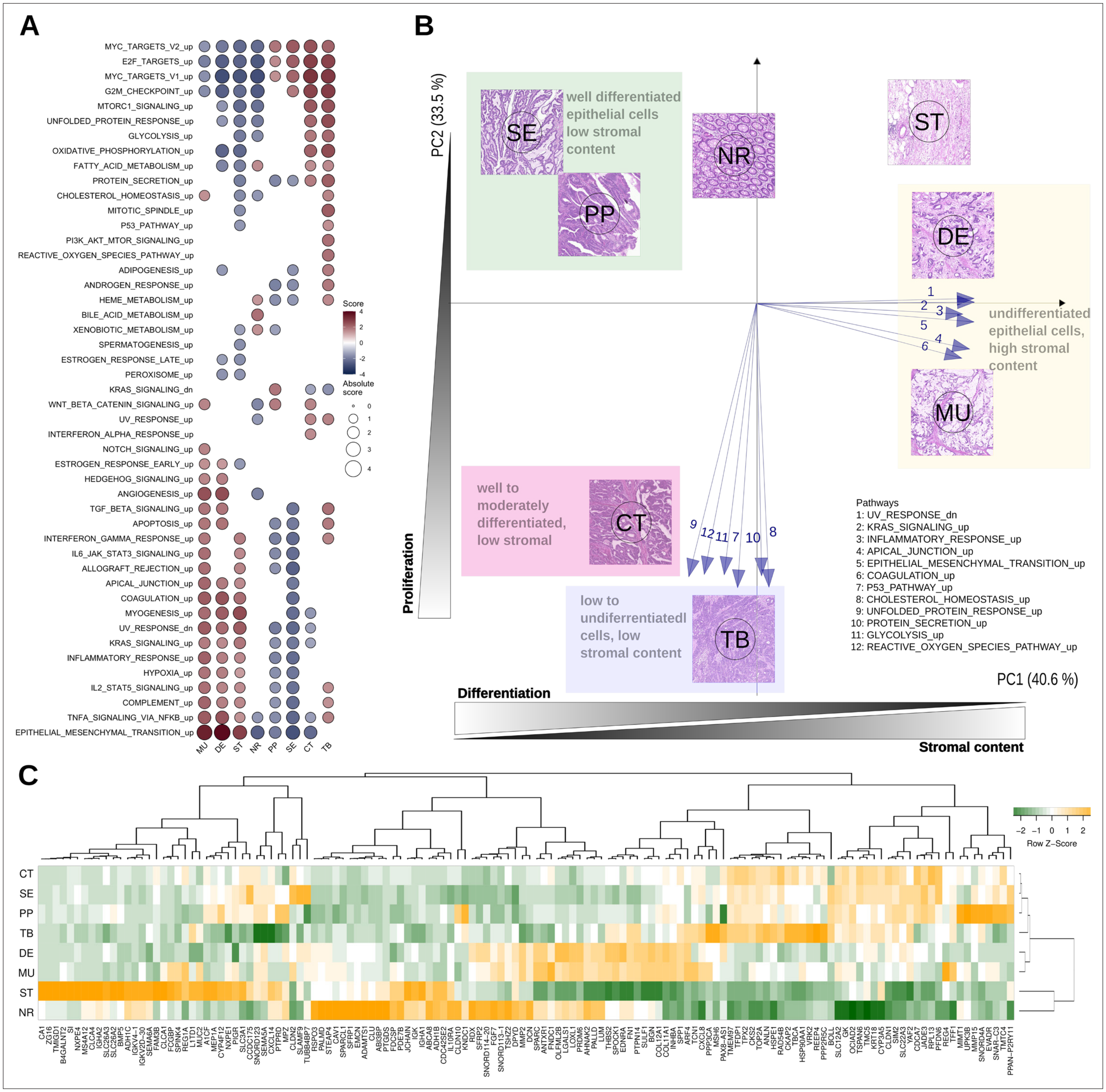


Figure 3. Top differentially expressed genes and hallmark pathways. (**A**) GSEA scores for hallmark pathways in the six morphotypes and two non-tumoral regions. Only pathways with statistically significant scores are shown. (**B**) Principal component analysis of hallmark pathways: the median profiles of the six morphotypes (CT: complex tubular, DE: desmoplastic, MU: mucinous, PP: papillary, SE: serrated, and TB: solid/trabecular) and the two non-tumoral regions (NR: tumor-adjacent normal and ST: supportive stroma) are projected onto the space defined by first two principal components (74% of the total variance). The top pathways contributing to the principal axes are shown as well. See also ***Figure 3—figure supplement 1***. (**C**) Heatmap of top 5 up- and down-regulated genes for each of the six morphotypes.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Principal component analysis of hallmark pathways GSEA scores: loadings for the first two principal components, i.e., contribution of pathways to the first two axes.

Figure supplement 2. Hallmark pathways differential activation between pairs of morphotypes.

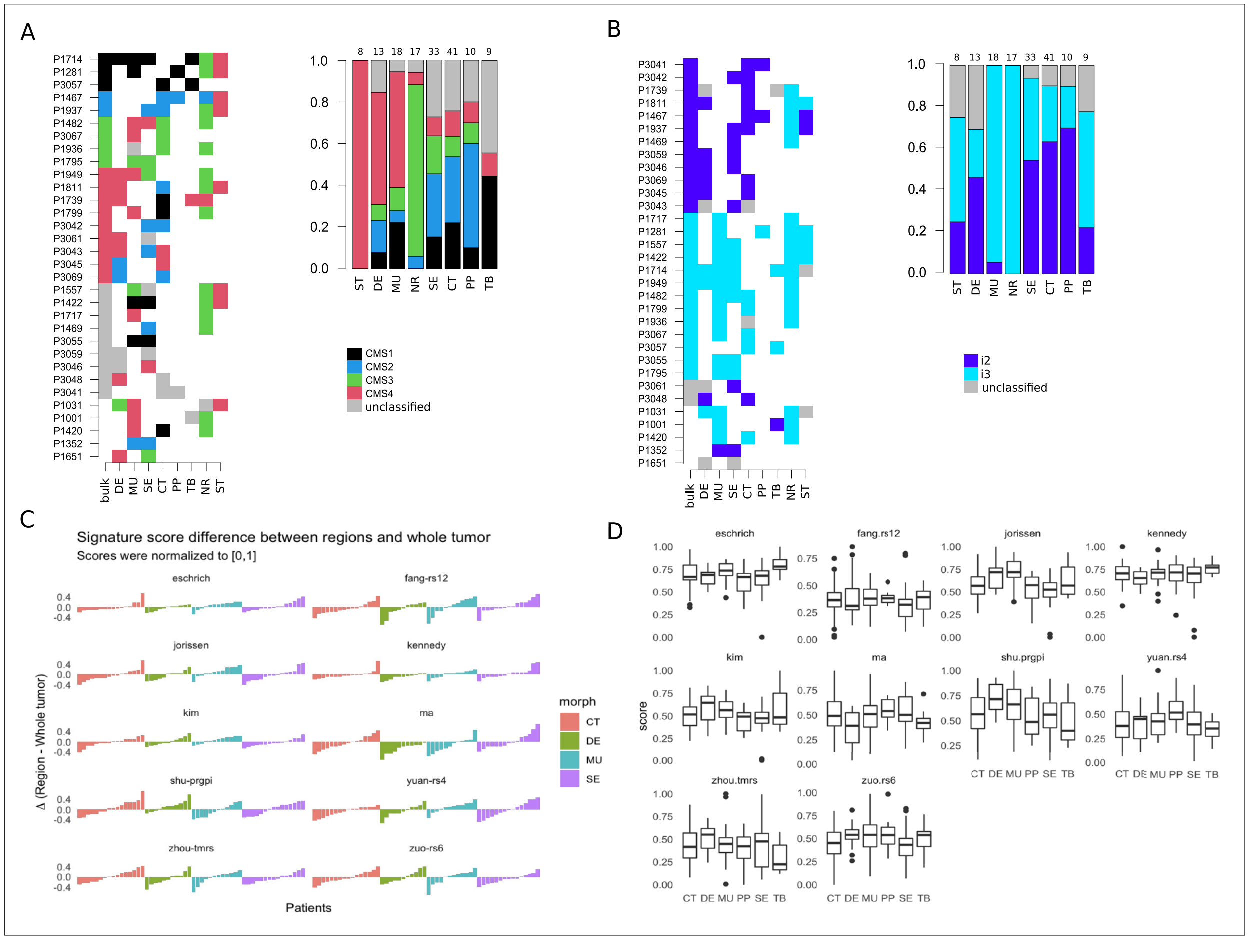


Figure 4. Intra-tumoral heterogeneity and the morphotypes (for all core samples, including those unassigned by the classifiers). Only cases with at least two distinct morphotypes present are shown. (**A**) Left: CMS assignment for tumors represented by multiple regions. Right: CMS assignment per morphotype (and two non-tumoral patterns). (**B**) Left: iCMS assignment for tumors represented by multiple regions. Right: iCMS assignment per morphotype (and two non-tumoral patterns). (**C**) Differences between paired signatures: morphotypes vs whole tumor (each signature was normalized to [0,1] prior to computing the differences). Only four (morphotype, whole tumor) pairs were represented enough in the data. (**D**) Boxplots for the ten (normalized) signatures across morphotypes. The 'Eschrich' and 'Jorissen' signatures vary significantly (Kruskal-Wallis's test) across morphotypes. For equivalent plots for all samples, including non-core, see ***Figure 4—figure supplement 1***.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Molecular subtypes and morphotypes in all samples, including non-core samples.

macrodissected morphological regions. We predicted both the consensus (CMS) (***Guinney et al., 2015***) and intrinsic (iCMS) (***Joanito et al., 2022***) molecular subtypes.

All ST regions were predicted as CMS4, and 82.4% of NR regions as CMS3. For the morphotypes, the predictions were more distributed across subtypes: DE and MU were most often assigned to CMS4 (63.6% and 58.8%), PP, SE, and CT to CMS2 (62.5%, 41.7% and 41.9%) and TB to CMS1 (80%; ***Figure 4A***, ***Figure 4—figure supplement 1***). More importantly, this heterogeneity was also observed intra-tumoral, with regions within the same tumor section being assigned to different subtypes (***Figures 4A and 5***, ***Figure 5—figure supplements 1 and 2***).

In contrast, intrinsic molecular subtypes (iCMS2/3) were much more stable, most of the time all the morphotypes within a tumor sharing the same iCMS label (***Figure 4B***, ***Figure 4—figure supplement 1***) and agreeing with the whole-tumor assignment. NR, MU, TB, and ST regions were classified most of the time as iCMS3 (100%, 94,4%, 71.4%, 66.7%), while PP, CT and DE were predominantly classified as iCMS2 (77.8%, 70.3%, 66.7%). The serrated morphotype was almost equally assigned to each of the iCMSs (iCMS2: 58%, iCMS3:42%).

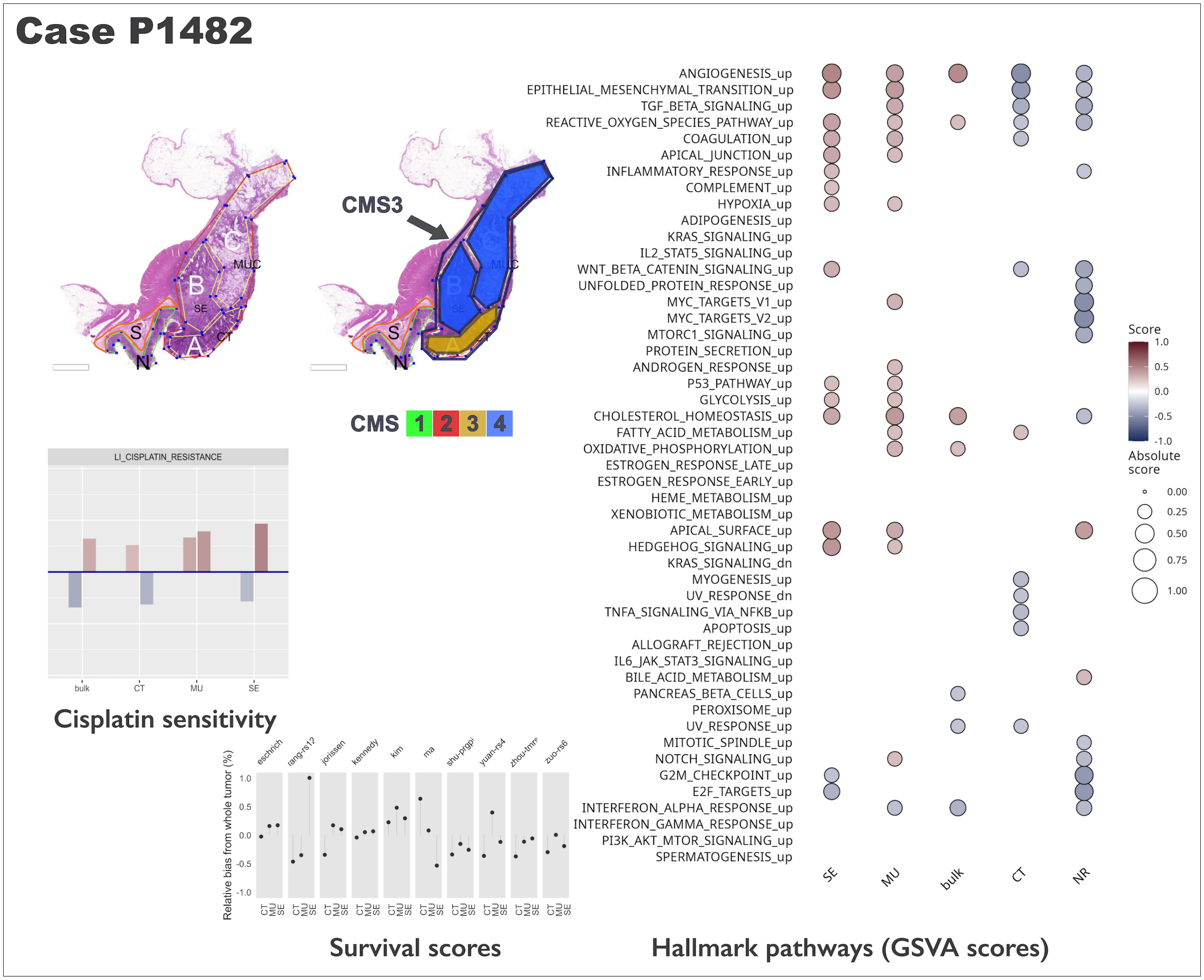


Figure 5. Intra-tumoral heterogeneity case study. For the same case, different CMS labels are assigned to regions and whole tumor profile. The hallmark pathways show various levels of activation (as computed by GSVA) within same section. The relative change in prognostic scores indicate potential underestimation of risk for some signatures, while others appear to be stable across tumor. See also ***Figure 5—figure supplements 1 and 2***. Note that in the pathology section image, the original annotations were preserved, and they are not identical to the ones used in the main text. Here, MUC stands for mucinous (MU) in the text. Also, N indicates a tumor-adjacent normal epithelial region and S a supportive stroma region, respectively.

The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Intra-tumoral heterogeneity additional case study.

Figure supplement 2. Intra-tumoral heterogeneity additional case study.

Prognostic and predictive gene-based signatures

The morphotypes generally differed in terms of score distributions, with two signatures reaching statistical significance (Kruskal-Wallis's test: Eschrich p=0.0228, Jorissen p=0.00085, ***Figure 4C–D***). A more pronounced variability was observed when comparing tumor regions to matched whole tumor, with amplitude of the differences (region vs whole tumor) larger than 50% of the whole tumor score in some cases (***Figure 4C***). ***Figure 5*** shows a case study with three different morphological regions (CT, MU, SE) which manifest rather large deviations from the whole tumor-based risk scores for most of the prognostic signatures (see also ***Figure 5—figure supplements 1 and 2***).

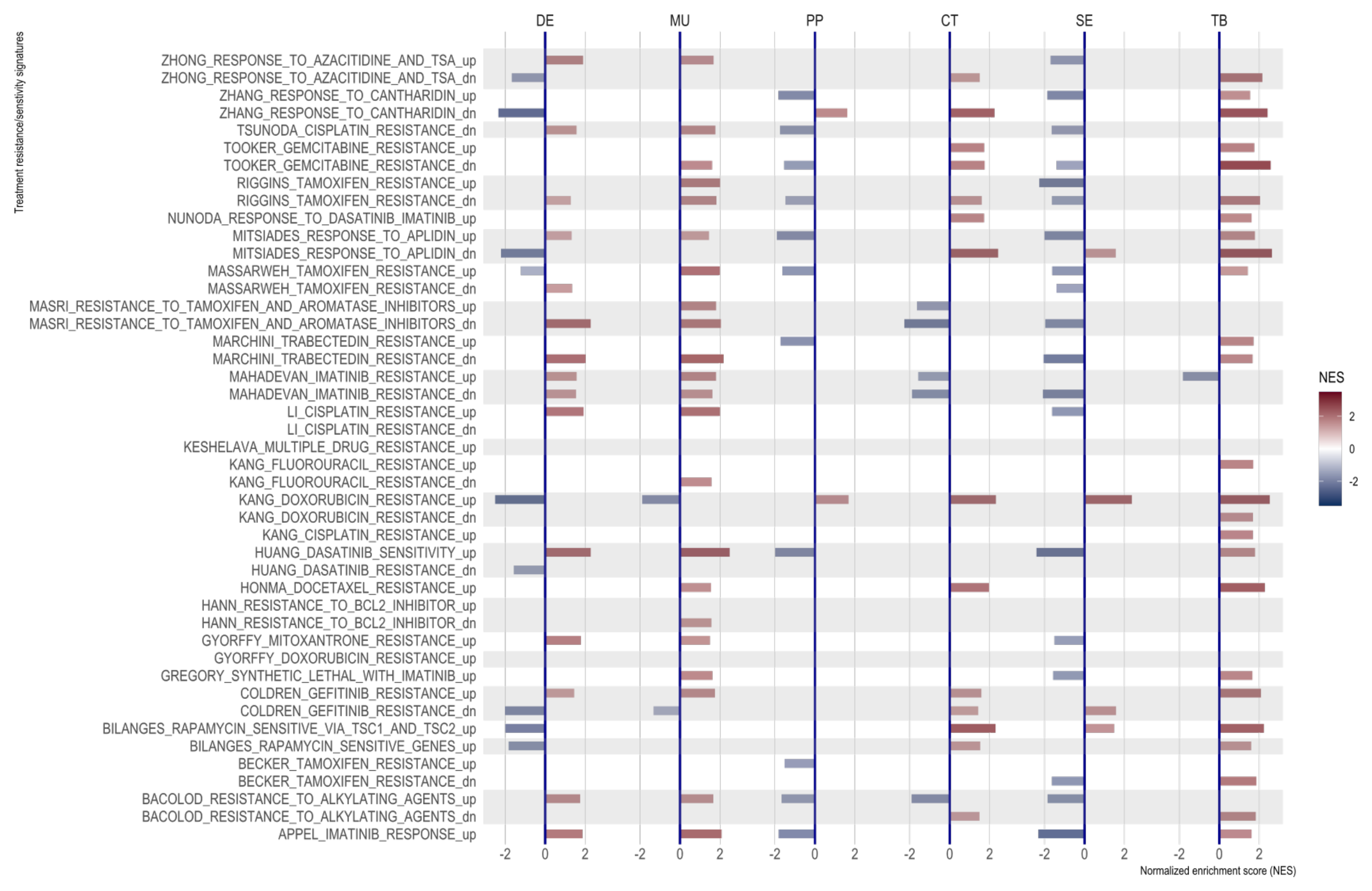


Figure 6. Normalized enrichment scores from GSEA for selected resistance signatures (from C2 section of MSigDB). Only significant scores are shown.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Resistance scores (GSVA) per patient and morphotype for cases where the whole–tumor prediction is contradicted by some regional score.

The predicted resistance/sensitivity to different therapeutics varied across morphotypes: MU resistance to gefitinib; DE sensitivity to azaticidine, dasatinib, and aplidin, and resistance to tamoxifen and gefitinib; PP resistance to cantharidin, SE resistance to aplidin, CT sensitivity to alkylating agents (***Figure 6***). The differences were observed even within tumor (***Figure 5***), with some of the supposedly sensitive tumors (whole tumor scoring) having regions of predicted resistance (***Figure 6—figure supplement 1***).

Discussion

The analysis of the morphotypes from transcriptomics perspective is meant to bridge the histopathology and gene expression. The present exploratory study was motivated by our earlier observations linking the morphological aspects of CRC to the molecular subtypes (***Budinska et al., 2013***). The original observations semi-quantitatively scored the morphotypes as primary or secondary dominant in the whole tumor section and showed that subtype A (corresponding to CMS3) was enriched in PP and SE morphologies, subtype B (corresponding to CMS2) in CT morphology, subtype C (corresponding to CMS1) in MU and TB morphologies, and, finally, subtype D (CMS4) in DE/stromal reaction (***Budinska et al., 2013***). In contrast, here we focused on tumor regions rather than whole tumor, which also allowed the characterization of the intra-tumor heterogeneity.

The results show a whole landscape of changes at gene and pathway levels, with morphotypes residing on a continuum space of molecular descriptors. The analysis of hallmark pathways and selected signatures combined with in silico deconvolution of cellular admixtures served two purposes. First, to confirm that the samples exhibit known properties (e.g. TB, SE, and PP have high tumor epithelial cell content, and DE and MU are enriched in fibroblasts; molecular EMT signature is high in MU and DE, but low in CT and SE, etc.), thus ensuring proper quality of the data. Second, it served to refine the characterization of the morphotypes and sketching their 'molecular portraits'. The morphotypes investigated had a fluid characterization from a transcriptomics perspective, with many pairwise similarities and some striking differences. Even from a strict histopathology perspective, it was difficult, if not impossible, to clearly distinguish the separation between adjacent morphological regions therefore a certain degree of contamination between morphotypes was to be expected. Nevertheless, the enrichment in specific cell types and states allowed the identification of characteristic molecular features.

MU and DE morphotypes (previously associated with CMS1 and CMS4, **Budinska et al., 2013**), as expected, exhibited high score of genes up-regulated in colon fibroblast TGF-β signaling pathway, genes associated with high tumor stromal content, CAFs and endothelial cells as well as pathways involved in immune system response. The detailed analysis of CAFs in fibroblast-rich regions (DE, MU, and ST) based on signatures derived from single cell sequencing studies (**Pelka et al., 2021**; **Kieffer et al., 2020**) revealed some finer differences: the supportive stroma (ST) region had the complete panel of fibroblast tested, while DE and MU most notably missed the 'normal CAFs'. The main difference between DE and MU appeared to be that former was enriched in CAFs associated with inflammatory response (IL-iCAF), all the other CAFs being present at similar levels. The other morphotypes were either significantly depleted in fibroblast signatures or their GSEA scores were not statistically significant. Deconvolution of immune cell fractions by quantiSeq showed enrichment of DE and MU in M1 macrophages. Given the involvement of CAFs in modeling the tumor microenvironment through ECM remodeling, angiogenesis promotion and immune system regulation (**Desbois and Wang, 2021**), our results support the idea of scoring separately the stromal component by either molecular or histopathology descriptors, in addition to tumor regions themselves. Even though DE and MU (and ST) also had the highest scores for the molecular EMT signature, our observations rather support the description of CMS4 as stromal/desmoplastic subtype than 'true' mesenchymal, in agreement with (**Loughrey et al., 2021**). Further, the poor prognostic associated with CMS4 could be explained by the stromal component: both (**Roseweir et al., 2020**) and (**Ten Hoorn et al., 2022**) agree that a high stromal invasion/desmoplastic reaction is prognostic of shorter time to relapse.

CT morphotype represents a classic adenocarcinoma and is one of the most common morphologies. In our previous study (**Budinska et al., 2013**), this morphotype was associated mainly with subtype B (vastly overlapping with CMS2). TB morphotype seems to be mostly representative of higher-grade tumors and was associated with CMS1. In contrast to NR, CT and TB showed significant enrichment of signatures of normal colon basal cells. From the molecular perspective, CT together with TB had both activated major pathways involved in proliferation processes. TB, in addition, resembled MU and DE morphotypes by sharing active TGF-β signaling, apoptosis, and active immune system response. SE and PP morphologies may be indicative of a different oncologic pathway – the 'serrated pathway' (**De Palma et al., 2019**). The two morphologies share common features like well to moderately differentiated, with low stromal content and crypt structure still preserved. From a molecular perspective, we found that both SE and PP were both distributed similarly across molecular subtypes (both CMS and iCMS) and had similar activation of hallmark pathways: EMT, *IL2/STAT5*, *IL6/STAT3*, *KRAS* signaling all being down-regulated, while *MYC* targets being up-regulated. Among the hallmark pathways, androgen response, heme metabolisms and *IL6/STAT3* (all silenced), appeared to be specific (and statistically significant) to SE and PP.

Given the relatively small sample size and similarities already observed between the morphotypes, it came as no surprise that the lists of differentially expressed genes, morphotype-specific, were generally short (for FDR ≤0.15). Nevertheless, literature search of the genes on top of these lists showed importance of these genes in CRC development, progression, EMT transition or response to therapy. For CT, the top gene was *PIP5K1B* which was related to PI3K/AKT signaling and seems to be involved in colorectal cancer development (**Zhang et al., 2019**). TB had the most differentially expressed genes (n=662) in comparison with all other morphotypes, with top genes including *FBXO5* – prognostic of shorter time-to-relapse in various cancers (**Liu et al., 2022**), *FLRT3* – a proapoptotic gene which, when overexpressed, inhibits EMT (**Yang et al., 2022**), *SETSIP* – gene coding chromatin-binding protein capable of participating in fibroblast reprogramming and differentiation into epithelial cells (**Margariti et al., 2012**), *E2F7* – up-regulated by *p53* in response to DNA damage (**Carvajal et al., 2012**), *CXCL14* (downregulated) - depending on the cell of origin can have both tumor suppressive or supporting role (**Westrich et al., 2020**), *SEMA5A* (downregulated) gene – proposed as prognostic marker in CRC (**Demirkol et al., 2017**). Among top overexpressed genes specific to MU morphotype we found *FGF7* (fibroblast growth factor 7) whose disrupted signaling was associated with deregulation of cell differentiation (**Patel et al., 2019**), and *MUC2* (intestinal mucin) whose downregulation has been suggested a marker of adverse outcome (**Betge et al., 2016**). At the same time, *MUC2* was also among the DE-specific genes, but downregulated, consistent with the observation that desmoplastic reaction is a marker of shorter relapse-free survival (**Ueno et al., 2021**). Still in DE, we found as top overexpressed genes *PIEZO2* – a paralog of *PIEZO1* which is involved in colorectal cancer metastases (**Sun et al., 2020**), *SLIT3* – a member of the Slit/Robo pathway, a major regulator of several oncogenic pathways and potential therapeutic target (**Gara et al., 2015**), and *OLFML2B* – a potential biomarker for resistance to MEK inhibitors (**Hu et al., 2022**). SE morphotype had only one specific gene overexpressed at FDR ≤0.15, *CCDC175*. At the other end of the list, very interestingly, we found significantly downregulated gene for dihydropyrimidine dehydrogenase (*DPYD*) gene – the variants of which are predictive of 5-fluoruracil toxicity in adjuvant colon cancer treatment (**Lee et al., 2014**), *GLIPR2* which participates in positive regulation of ERK1/2 cascade and EMT transition (**Kang et al., 2012b**), or the *HOXA9A* gene, the overexpression of which was suggested to contribute to stem cell overpopulation responsible for development of CRC (**Osmond et al., 2022**) or the *GLI3* gene – that participates in sonic hedgehog (Shh)-Gli-mediated tumorigenesis and the loss of Gli3 signaling was shown to initiate cell growth inhibition in colon cancer cells, while sensitizing colon cancer cells to treatment with anticancer agents (5-FU and bevacizumab) (**Kang et al., 2012a**). The only specific gene marker of the PP morphotype was the downregulation of *MZP* – myelin protein zero.

We also found significant differences between pairs of morphotypes, especially in terms of molecular signatures/programs. These results reinforce the observations above and show that they are robust to the proportion of fibroblasts and/or epithelial cells present in the compared morphotypes.

In our collection, several cases were represented by several regions and an additional whole-tumor profile. Taking advantage of these matched samples, we investigated several molecular classifiers from an intra-tumor variability perspective as well. The CMS classification was less stable than iCMS, with whole tumor CMS class differing from at least one of the constituent morphological regions in about 60% of cases (11 out of 18, excluding cases in which CMS class was not predicted; see **Figure 5**). Additionally, we tested several prognostic and predictive expression-based classifiers/signatures. The goal was not to compare them in terms of their predictive capabilities (the experimental design did not allow for such an exercise), but rather to have a clear picture of the extent to which the various morphotypes 'distract' these predictors. We found that all the prognostic signatures varied with the morphological regions with some striking cases in which the morphotype scores exceeded the corresponding whole tumor scores by more than 50%. This observation suggests that, in some cases, the whole tumor-based predictions were too optimistic, the models failing to recognize higher risk cases. While these signatures were derived from whole-tumor expression profiles, their variability across tumor indicates the need for precise tumor sampling strategies.

Our exploratory study has, inherently, several limitations. The selection of cases may not represent the proportions of various morphotypes found in general population of CRC patients. Our selection tried to cover as many scenarios as feasible with a limited number of samples. Also, the tumor heterogeneity in terms of morphotypes cannot be estimated from these data since a single tissue block per tumor was considered. The reduced sample size in some of the paired comparisons within same tumor calls for further external validation. However, our results pave the way to future studies addressing these questions and others related to optimizing the tumor sampling strategy, for example.

We have analyzed the gene expression profiles of six morphotypes (and two peritumoral regions), building a comprehensive molecular picture of their salient features. The observed heterogeneity, especially intra-tumoral (**Figure 5**), calls for a finer resolution of the tumor sampling in profiling studies. Until spatial transcriptomics becomes integrated in routine clinical practice, using the morphotypes for anchoring the expression profiles is a feasible approach. Our study already provides indications of the molecular programs one would expect to find de-/activated in these regions, thus helping in designing future experiments. The implications for molecular classifiers are clear: it is necessary to account for tumor morphology when designing new biomarkers. Given the sensitivity of many gene-based classifiers to the tumor and stroma proportions in the samples, there is a need to adjust these classifiers to control for their relative proportions. This can be achieved by different means, and we presented an approach based on morphotypes.

From a molecular pathology practice perspective, the molecular descriptors found to vary across morphotypes may help in patient stratification and provide hints for further, more targeted investigations. Several questions call for further investigation: (i) how much of a tumor needs to be embedded to achieve a precise molecular diagnostic? and (ii) what precise tumor region(s) are needed for a molecular diagnostic? The morphotypes selected here may need further refinement and achieving consensus among pathologists regarding their exact definition, a point that could potentially be addressed by automatic image analysis approaches.

Ideas and speculation

Our analyses indicate that both prognostic and response to therapy signatures may predict more severe cases (shorter relapse free survival or resistance to therapy) when applied to subregions than to the whole tumor. This might be one of the reasons the said signatures may fail their real-world validation. Therefore, morphologically heterogeneous tumors need several sampling locations to provide a more sensible result. Sensitivity and cost analyses need to be performed to estimate the benefits of multi-regional sampling.

Further, the fact that we were able to identify specific molecular programs associated with the morphotypes calls for investigating the inverse problem as well, that is whether sufficiently discriminatory features could be extracted for estimating the proportions of the morphotypes from whole tumor profiles.

Materials and methods

Samples

This retrospective cross-sectional study used tumor samples from patients with CRC who were examined at Masaryk Memorial Cancer Institute, Brno, Czech Republic in years 2002–2015. The study was reviewed and approved by the Committee for Ethics of Masaryk Memorial Cancer Institute, Brno, Czech Republic (number 2018/861/MOU). All patients gave written informed consent for the use of their biological samples for research purposes. Fundamental ethical principles and rights promoted by the European Union EU (2000/C364/01) were followed. All patients' data were processed according to the Declaration of Helsinki (last revision 2013). Inclusion criteria for this study were: age >18 years, clinical and histopathologically confirmed diagnosis of primary CRC. Standard clinical and histopathological variables (TNM, grade etc.) were retrieved for all patients. Failure of laboratory analyses (problematic sample preparation, low quality and/or quantity of isolated RNA, low quality of expression data) was a reason for excluding these samples from the study.

Sample preparation

A total of 111 colon cancers (unique patients) were identified in the tumor archive of the Masaryk Memorial Cancer Institute and were assessed by two expert pathologists. Morphological regions of interest, representing complex tubular (CT), desmoplastic (DE), mucinous (MU), papillary (PP), serrated (SE) and solid/trabecular (TB) morphologies, respectively (see ***Figure 1***), were digitally marked in scanned whole slide images (at 20 x magnification) and macrodissected for RNA extraction. Additionally, from several slides, tumor-adjacent normal (NR) and tumor-associated stroma (ST). Tumor samples with limited contamination of additional morphologies (<20%) were called 'core samples' and used morphotype molecular characterization. The labelling of the regions was repeated after 1 year to ensure a stable assignment. For n=28 cases, whole-tumor regions were macrodissected from the histology section immediately adjacent to the section used for morphological regions. Standard clinical and histopathological variables were retrieved for most of the patients.

Gene expression profiling

The RNA extraction was performed from formalin-fixed paraffin-embedded histopathological slides using AllPrep DNA/RNA Kits (Qiagen, Hilden, Germany) according to their specific manufacturer's instructions. A few modifications were made to the protocol: FFPE slides (2x3 µm) were bathed in a solution to remove paraffin (3 x in xylene for 5 min and 3 x in ethanol for 5 min). Tumor tissue was spotted with 8 ul PKD puffer and collected from slides using a scalpel. Purification was done for total RNA, including small RNAs. For elution, 20 ul RNA free water (1 min. incubation) was used and then repeated with eluate. The extracted RNA served as input for a GeneChip WT Pico Reagent Kit (Thermo Fisher Scientific, Waltham, MA, USA) for analysis of the transcriptome on whole-transcriptome arrays. We selected the input amount from the recommended range according to the manufacturer's instructions. Total RNA from HeLa cells provided in the kit was used as a positive control together with a high-quality low-concentration RNA isolated from a serum as a low input control. Clariom D Array for human samples (Thermo Fisher Scientific, Waltham, MA, USA) was used for target hybridization to capture both coding and multiple forms of non-coding RNA. Finally, the arrays were scanned using Affymetrix GeneChip Scanner 3000 7 G (Thermo Fisher Scientific, Waltham, MA, USA). The sample preparation and analysis were performed according to the manufacturer's instructions. The protocol included several control points in which the workflow was monitored. All the samples complied with the quality control requirements and none of the samples were excluded from the analysis.

The data generated in this study are publicly available in ArrayExpress under accession number E-MTAB-12599 (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-12599).

Bioinformatics analyses

All resulting CEL files were processed using Bioconductor (RRID:SCR_006442) (***Huber et al., 2015***) (v.3.15) packages oligo (***Carvalho and Irizarry, 2010***) (v.1.60), affycoretools (v1.68) and, for Clariom D chip annotation, pd.clariom.d.human (v.3.14). For the quality control we used AffyPLM (v.147) and imposed a maximal median Normalized Unscaled Standard Errors (NUSE) of 1.12. In all, n=202 passed all the quality control steps and were normalized together using RMA (oligo) with core-probeset summarization. Further, the array data was summarized at gene level by selecting the most variable probeset per unique EntrezID and entries corresponding to missing HUGO symbols, speculative transcripts, and short non-coding RNA were discarded resulting in a reduced list of 27,302 unique genes. Batch effects were removed using ComBat (***Johnson et al., 2007***) from package sva (v.3.44.0).

For the identification of differentially expressed genes we used linear models (limma package v.3.52.2) with a cut-off for false discovery rate FDR = 0.15. The pathways were scored in terms of enrichment in specific signatures using gene set enrichment analysis (GSEA) (***Subramanian et al., 2005***) as implemented in fgsea package (v.1.22.0). For scoring the signatures in individual samples, we used gene score variation analysis (GSVA) (***Hänzelmann et al., 2013***) implemented in GSVA package (v.1.44.1). MSigDB (RRID:SCR_016863) (all collections: H, C1-8; v.7.4.1) (***Liberzon et al., 2015***) was used as the main source for gene sets and pathways. Additional cell type-specific gene sets, some derived from whole tumor others from single-cell sequencing studies, representing (i) cancer associated fibroblasts (CAFs) (***Isella et al., 2015***; ***Pelka et al., 2021***; ***Khaliq et al., 2022***; ***Kieffer et al., 2020***) (ii) epithelial cells (***Kosinski et al., 2007***; ***Merlos-Suárez et al., 2011***; ***Pelka et al., 2021***), and (iii) immune cells (***Isella et al., 2015***; ***Pelka et al., 2021***) were used (see ***Supplementary file 3*** for full list). The consensus molecular subtypes were predicted using CMSCaller (***Eide et al., 2017***) (v.2.0.1) and the intrinsic epithelial subtypes (***Joanito et al., 2022***) using the signatures therein (P. Tsantoulis, personal communication, July 2022). The cellular mixture of various tumoral regions was explored computationally using quanTIseq (***Finotello et al., 2019***) (for immune cells) and ESTIMATE (***Yoshihara et al., 2013***) (for tumor purity/epithelial cells). The core samples were used for deriving the lists of differentially expressed genes, for gene set enrichment analyses and for in silico deconvolutions of cell populations. The analyses treating the samples independently were applied to all samples, including non-core.

Ten different survival/prognostic genomic signatures (full list in ***Supplementary file 13***) were computed per-sample as (weighted, when weights were provided) means of signature genes, and 29 sensitivity/resistance signatures selected from MSigDB/C2 were scored by GSVA.

All data analyses were performed in R 4.2 (***R Development Core Team, 2022***).

Acknowledgements

We thank Dr. Petros Tsantoulis, Hôpitaux Universitaires de Genève, Switzerland, for providing us with a model for classification of intrinsic epithelial subtypes based on gene expression data. The authors acknowledge funding from the Czech Science Foundation (GACR) grant no. GA19-08646S. Also, the support of the Research Infrastructure RECETOX RI (No LM2018121) and Cetocoen Plus project (CZ.02.1.01/0.0/0.0/15_003/0000469) financed by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS), the Teaming project (CETOCOEN Excellence 857560; CZ.02.1.01/0.0/0.0/17_043/0009632), and the project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union - Next Generation EU is acknowledged. This work was supported from the European Union's Horizon 2020 research and innovation program under grant agreement No 857560. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Additional information

Funding

Funder	Grant reference number	Author
Grantova Agentura Ceske Republiky	GA19-08646S	Eva Budinská Martina Hrivňáková Vlad Popovici

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Eva Budinská, Conceptualization, Data curation, Formal analysis, Validation, Visualization, Methodology, Writing – original draft, Project administration, Writing – review and editing; Martina Hrivňáková, Data curation, Methodology, Writing – review and editing; Tina Catela Ivkovic, Lenka Zdražilová Dubská, Data curation, Validation, Writing – original draft, Writing – review and editing; Marie Madrzyk, Data curation, Software, Writing – original draft; Rudolf Nenutil, Resources, Methodology, Writing – original draft; Beatrix Bencsiková, Data curation, Formal analysis, Writing – original draft, Writing – review and editing; Dagmar Al Tukmachi, Data curation, Methodology, Writing – original draft, Writing – review and editing; Michaela Ručková, Software, Validation, Writing – original draft; Ondřej Slabý, Supervision, Investigation, Methodology, Writing – original draft; Josef Feit, Supervision, Validation, Investigation, Writing – original draft; Mihnea-Paul Dragomir, Formal analysis, Investigation, Writing – original draft, Writing – review and editing; Petra Borilova Linhartova, Investigation, Writing – original draft, Writing – review and editing; Sabine Tejpar, Supervision, Investigation, Writing – original draft; Vlad Popovici, Conceptualization, Supervision, Funding acquisition, Investigation, Methodology, Writing – original draft, Project administration, Writing – review and editing

Author ORCIDs

Vlad Popovici http://orcid.org/0000-0002-1311-9188

Ethics

Human subjects: This retrospective cross-sectional study used tumor samples from patients with CRC who were examined at Masaryk Memorial Cancer Institute, Brno, Czech Republic in years 2002-2015. The study was reviewed and approved by the Committee for Ethics of Masaryk Memorial Cancer Institute, Brno, Czech Republic (number 2018/861/MOU). All patients gave written informed consent for the use of their biological samples for research purposes. Fundamental ethical principles and rights promoted by the European Union EU (2000/C364/01) were followed. All patients' data were processed according to the Declaration of Helsinki (last revision 2013). Inclusion criteria for this study were: age > 18 years, clinical and histopathologically confirmed diagnosis of primary CRC. Standard clinical and histopathological variables (TNM, grade etc.) were retrieved for all patients. Failure of laboratory analyses (problematic sample preparation, low quality and/or quantity of isolated RNA, low quality of expression data) was a reason for excluding these samples from the study.

Peer review material

Joint Public Review: https://doi.org/10.7554/eLife.86655.3.sa1
Author Response https://doi.org/10.7554/eLife.86655.3.sa2

Additional files

Supplementary files

- Supplementary file 1. Main clinical parameters of the study cohort.
- Supplementary file 2. Distribution of main clinical parameters per morphotype (and tumor-adjacent normal and supportive stroma).
- Supplementary file 3. Table of gene expression signatures.
- Supplementary file 4. Table of GSEA scores (NES) for "other" signatures (see also ***Supplementary file 3*** for signatures).
- Supplementary file 5. List of differentially expressed genes (limma tables) per morphotype in contrast with pooled profile.
- Supplementary file 6. GSEA results for genes in ***Supplementary file 5***, for whole MSigDB collection.
- Supplementary file 7. List of differentially expressed genes (limma tables) per morphotype in contrast with all other five morphotypes.
- Supplementary file 8. GSEA results for genes in ***Supplementary file 7***, for whole MSigDB collection.
- Supplementary file 9. List of differentially expressed genes (limma tables) per matched pairs of morphotypes.
- Supplementary file 10. GSEA results for genes in ***Supplementary file 9***, for whole MSigDB collection.
- Supplementary file 11. List of differentially expressed genes (limma tables) for pairs of morphotypes.
- Supplementary file 12. GSEA results for genes in ***Supplementary file 11***, for whole MSigDB collection.
- Supplementary file 13. List of prognostic signatures tested.
- MDAR checklist

Data availability

Data generated through this study is publicly available from ArrayExpress under accession number E-MTAB-12599. All analyses results are available as supplementary files.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Popovici V, Budinska E	2023	Molecular portraits of colorectal tumor morphological regions	https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-12599	ArrayExpress, E-MTAB-12599

References

Amin MB. 2017. American Cancer society. Amin MB, Edge SB, Gress DM, Meyer LR (Eds). *AJCC Cancer Staging Manual* Chicago IL: Springer. p. XVII–1032.

Betge J, Schneider NI, Harbaum L, Pollheimer MJ, Lindtner RA, Kornprat P, Ebert MP, Langner C. 2016. MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: expression profiles and clinical significance. *Virchows Archiv* **469**:255–265. DOI: https://doi.org/10.1007/s00428-016-1970-5, PMID: 27298226

Bosman FT. 2010. *WHO Classification of Tumours of the Digestive System* Lyon: International Agency for Research on Cancer.

Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO. 2013. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology* **231**:63–76. DOI: https://doi.org/10.1002/path.4212

Budinska E, Bosman F, Popovici V. 2016. Experiments in molecular subtype recognition based on histopathology images. 2016 IEEE 13th International Symposium on Biomedical Imaging. 1168–1172. DOI: https://doi.org/10.1109/ISBI.2016.7493474

Carvajal LA, Hamard PJ, Tonnessen C, Manfredi JJ. 2012. E2F7, a novel target, is up-regulated by p53 and mediates DNA damage-dependent transcriptional repression. *Genes & Development* **26**:1533–1545. DOI: https://doi.org/10.1101/gad.184911.111, PMID: 22802528

Carvalho BS, Irizarry RA. 2010. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**:2363–2367. DOI: https://doi.org/10.1093/bioinformatics/btq431, PMID: 20688976

Demirkol S, Gomceli I, Isbilen M, Dayanc BE, Tez M, Bostanci EB, Turhan N, Akoglu M, Ozyerli E, Durdu S, Konu O, Nissan A, Gonen M, Gure AO. 2017. A combined ULBP2 and SEMA5A expression signature as a prognostic and predictive Biomarker for Colon Cancer. *Journal of Cancer* **8**:1113–1122. DOI: https://doi.org/10.7150/jca.17872, PMID: 28607584

De Palma FDE, D'Argenio V, Pol J, Kroemer G, Maiuri MC, Salvatore F. 2019. The Molecular Hallmarks of the serrated pathway in Colorectal Cancer. *Cancers* **11**:1017. DOI: https://doi.org/10.3390/cancers11071017, PMID: 31330830

Desbois M, Wang Y. 2021. Cancer-associated fibroblasts: Key players in shaping the tumor immune microenvironment. *Immunological Reviews* **302**:241–258. DOI: https://doi.org/10.1111/imr.12982, PMID: 34075584

De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij L, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, Rodermond H, van der Heijden M, van Noesel CJM, Tuynman JB, Dekker E, Markowetz F, Medema JP, Vermeulen L. 2013. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine* **19**:614–618. DOI: https://doi.org/10.1038/nm.3174, PMID: 23584090

Dunne PD, McArt DG, Bradley CA, O'Reilly PG, Barrett HL, Cummins R, O'Grady T, Arthur K, Loughrey MB, Allen WL, McDade SS, Waugh DJ, Hamilton PW, Longley DB, Kay EW, Johnston PG, Lawler M, Salto-Tellez M, Van Schaeybroeck S. 2016. Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. *Clinical Cancer Research* **22**:4095–4104. DOI: https://doi.org/10.1158/1078-0432.CCR-16-0032, PMID: 27151745

Eide PW, Bruun J, Lothe RA, Sveen A. 2017. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Scientific Reports* **7**:16618. DOI: https://doi.org/10.1038/s41598-017-16747-x, PMID: 29192179

Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, Sopper S, Ijsselsteijn M, Brouwer TP, Johnson D, Xu Y, Wang Y, Sanders ME, Estrada MV, Ericsson-Gonzalez P, Charoentong P, et al. 2019. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Medicine* **11**:34. DOI: https://doi.org/10.1186/s13073-019-0638-6, PMID: 31358023

Gara RK, Kumari S, Ganju A, Yallapu MM, Jaggi M, Chauhan SC. 2015. Slit/Robo pathway: a promising therapeutic target for cancer. *Drug Discovery Today* **20**:156–164. DOI: https://doi.org/10.1016/j.drudis.2014.09.008, PMID: 25245168

Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, et al. 2015. The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**:1350–1356. DOI: https://doi.org/10.1038/nm.3967, PMID: 26457759

Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**:7. DOI: https://doi.org/10.1186/1471-2105-14-7, PMID: 23323831

Hu P, Zhang X, Li Y, Xu L, Qiu H. 2022. Pan-Cancer analysis of OLFML2B expression and its association with Prognosis and Immune Infiltration. *Frontiers in Genetics* **13**:882794. DOI: https://doi.org/10.3389/fgene.2022.882794, PMID: 35873458

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**:115–121. DOI: https://doi.org/10.1038/nmeth.3252, PMID: 25633503

Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, Mellano A, Senetta R, Cassenti A, Sonetto C, Inghirami G, Trusolino L, Fekete Z, De Ridder M, Cassoni P, Storme G, Bertotti A, Medico E. 2015. Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics* **47**:312–319. DOI: https://doi.org/10.1038/ng.3224, PMID: 25706627

Jass JR. 2007. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **50**:113–130. DOI: https://doi.org/10.1111/j.1365-2559.2006.02549.x, PMID: 17204026

Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, Eng CLP, Macalinao DC, Kahraman M, Srinivasan H, Lakshmanan V, Verbandt S, Tsantoulis P, Gunn N, Venkatesh PN, Poh ZW, Nahar R, Oh HLJ, Loo JM, Chia S, et al. 2022. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics* **54**:963–975. DOI: https://doi.org/10.1038/s41588-022-01100-4, PMID: 35773407

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**:118–127. DOI: https://doi.org/10.1093/biostatistics/kxj037, PMID: 16632515

Kang J, D'Andrea AD, Kozono D. 2012a. A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *Journal of the National Cancer Institute* **104**:670–681. DOI: https://doi.org/10.1093/jnci/djs177, PMID: 22505474

Kang HN, Oh SC, Kim JS, Yoo YA. 2012b. Abrogation of Gli3 expression suppresses the growth of colon cancer cells via activation of p53. *Experimental Cell Research* **318**:539–549. DOI: https://doi.org/10.1016/j.yexcr.2011.12.010, PMID: 22227409

Khaliq AM, Erdogan C, Kurt Z, Turgut SS, Grunvald MW, Rand T, Khare S, Borgia JA, Hayden DM, Pappas SG, Govekar HR, Kam AE, Reiser J, Turaga K, Radovich M, Zang Y, Qiu Y, Liu Y, Fishel ML, Turk A, et al. 2022. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biology* **23**:113. DOI: https://doi.org/10.1186/s13059-022-02677-z, PMID: 35831907

Kieffer Y, Hocine HR, Gentric G, Pelon F, Bernard C, Bourachot B, Lameiras S, Albergante L, Bonneau C, Guyard A, Tarte K, Zinovyev A, Baulande S, Zalcman G, Vincent-Salomon A, Mechta-Grigoriou F. 2020. Single-Cell Analysis Reveals Fibroblast Clusters Linked to Immunotherapy Resistance in Cancer. *Cancer Discovery* **10**:1330–1351. DOI: https://doi.org/10.1158/2159-8290.CD-19-1384, PMID: 32434947

Kosinski C, Li VSW, Chan ASY, Zhang J, Ho C, Tsui WY, Chan TL, Mifflin RC, Powell DW, Yuen ST, Leung SY, Chen X. 2007. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *PNAS* **104**:15418–15423. DOI: https://doi.org/10.1073/pnas.0707210104, PMID: 17881565

Lee AM, Shi Q, Pavey E, Alberts SR, Sargent DJ, Sinicrope FA, Berenberg JL, Goldberg RM, Diasio RB. 2014. DPYD variants as predictors of 5-fluorouracil toxicity in adjuvant colon cancer treatment (NCCTG N0147). *Journal of the National Cancer Institute* **106**:dju298. DOI: https://doi.org/10.1093/jnci/dju298, PMID: 25381393

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems* **1**:417–425. DOI: https://doi.org/10.1016/j.cels.2015.12.004, PMID: 26771021

Liu P, Wang X, Pan L, Han B, He Z. 2022. Prognostic significance and immunological role of FBXO5 in Human Cancers: a systematic pan-cancer analysis. *Frontiers in Immunology* **13**:901784. DOI: https://doi.org/10.3389/fimmu.2022.901784, PMID: 35720327

Loughrey MB, Fisher NC, McCooey AJ, Dunne PD. 2021. Comment on "Identification of EMT-related high-risk stage II colorectal cancer and characterisation of metastasis-related genes *British Journal of Cancer* **124**:1175–1176. DOI: https://doi.org/10.1038/s41416-020-01213-9, PMID: 33311590

Margariti A, Winkler B, Karamariti E, Zampetaki A, Tsai T, Baban D, Ragoussis J, Huang Y, Han J-DJ, Zeng L, Hu Y, Xu Q. 2012. Direct reprogramming of fibroblasts into endothelial cells capable of angiogenesis and reendothelialization in tissue-engineered vessels. *PNAS* **109**:13793–13798. DOI: https://doi.org/10.1073/pnas.1205526109, PMID: 22869753

Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, et al. 2013. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLOS Medicine* **10**:e1001453. DOI: https://doi.org/10.1371/journal.pmed.1001453, PMID: 23700391

Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, Rossell D, Sevillano M, Hernando-Momblona X, da Silva-Diz V, Muñoz P, Clevers H, Sancho E, Mangues R, Batlle E. 2011. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**:511–524. DOI: https://doi.org/10.1016/j.stem.2011.02.020, PMID: 21419747

Müller MF, Ibrahim AEK, Arends MJ. 2016. Molecular pathological classification of colorectal cancer. *Virchows Archiv* **469**:125–134. DOI: https://doi.org/10.1007/s00428-016-1956-3, PMID: 27325016

Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**:330–337. DOI: https://doi.org/10.1038/nature11252, PMID: 22810696

Osmond B, Facey COB, Zhang C, Boman BM. 2022. HOXA9 Overexpression Contributes to Stem Cell Overpopulation That Drives Development and Growth of Colorectal Cancer. *International Journal of Molecular Sciences* **23**:6799. DOI: https://doi.org/10.3390/ijms23126799, PMID: 35743243

Patel A, Tripathi G, McTernan P, Gopalakrishnan K, Ali O, Spector E, Williams N, Arasaradnam RP. 2019. Fibroblast growth factor 7 signalling is disrupted in colorectal cancer and is a potential marker of field cancerisation. *Journal of Gastrointestinal Oncology* **10**:429–436. DOI: https://doi.org/10.21037/jgo.2019.02.11, PMID: 31183192

Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, Bejnood A, Dionne D, Ge WH, Xu KH, Chao SX, Zollinger DR, Lieb DJ, Reeves JW, Fuhrman CA, Hoang ML, Delorey T, Nguyen LT, Waldman J, Klapholz M, et al. 2021. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**:4734–4752.. DOI: https://doi.org/10.1016/j.cell.2021.08.003, PMID: 34450029

Perez-Villamil B, Romera-Lopez A, Hernandez-Prieto S, Lopez-Campos G, Calles A, Lopez-Asenjo JA, Sanz-Ortega J, Fernandez-Perez C, Sastre J, Alfonso R, Caldes T, Martin-Sanchez F, Diaz-Rubio E. 2012. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer* **12**:260. DOI: https://doi.org/10.1186/1471-2407-12-260, PMID: 22712570

Popovici V, Budinská E, Dušek L, Kozubek M, Bosman F. 2017. Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. *Bioinformatics* **33**:2002–2009. DOI: https://doi.org/10.1093/bioinformatics/btx027, PMID: 28158480

Rao A, Barkley D, França GS, Yanai I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**:211–220. DOI: https://doi.org/10.1038/s41586-021-03634-9, PMID: 34381231

R Development Core Team. 2022. R: A language and environment for statistical computing. 2.6.2. Vienna, Austria. R Foundation for Statistical Computing. https://www.r-project.org

Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, Snel MH, Chresta CM, Rosenberg R, Nitsche U, Macarulla T, Capella G, Salazar R, Orphanides G, Wessels LFA, Bernards R, Simon IM. 2014. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer* **134**:552–562. DOI: https://doi.org/10.1002/ijc.28387, PMID: 23852808

Roseweir AK, Park JH, Hoorn ST, Powell AG, Aherne S, Roxburgh CS, McMillan DC, Horgan PG, Ryan E, Sheahan K, Vermeulen L, Paul J, Harkin A, Graham J, Sansom O, Church DN, Tomlinson I, Saunders M, Iveson TJ, Edwards J. 2020. Histological phenotypic subtypes predict recurrence risk and response to adjuvant chemotherapy in patients with stage III colorectal cancer. *The Journal of Pathology. Clinical Research* **6**:283–296. DOI: https://doi.org/10.1002/cjp2.171, PMID: 32401426

Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LCG, Lannon WA, Grotzinger C, Del Rio M, Lhermitte B, Olshen AB, Wiedenmann B, Cantley LC, Gray JW, Hanahan D. 2013. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine* **19**:619–625. DOI: https://doi.org/10.1038/nm.3175, PMID: 23584089

Stewart JP, Richman S, Maughan T, Lawler M, Dunne PD, Salto-Tellez M. 2017. Standardising RNA profiling based biomarker application in cancer-The need for robust control of technical variables. *Biochimica et Biophysica Acta. Reviews on Cancer* **1868**:258–272. DOI: https://doi.org/10.1016/j.bbcan.2017.05.005, PMID: 28549623

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**:15545–15550. DOI: https://doi.org/10.1073/pnas.0506580102, PMID: 16199517

Sun Y, Li M, Liu G, Zhang X, Zhi L, Zhao J, Wang G. 2020. The function of Piezo1 in colon cancer metastasis and its potential regulatory mechanism. *Journal of Cancer Research and Clinical Oncology* **146**:1139–1152. DOI: https://doi.org/10.1007/s00432-020-03179-w, PMID: 32152662

Tang X, Huang Y, Lei J, Luo H, Zhu X. 2019. The single-cell sequencing: new developments and medical applications. *Cell & Bioscience* **9**:53. DOI: https://doi.org/10.1186/s13578-019-0314-y, PMID: 31391919

Ten Hoorn S, de Back TR, Sommeijer DW, Vermeulen L. 2022. Clinical value of consensus Molecular Subtypes in Colorectal Cancer: a systematic review and meta-analysis. *Journal of the National Cancer Institute* **114**:503–516. DOI: https://doi.org/10.1093/jnci/djab106, PMID: 34077519

Ueno H, Ishiguro M, Nakatani E, Ishikawa T, Uetake H, Murotani K, Matsui S, Teramukai S, Sugai T, Ajioka Y, Maruo H, Kotaka M, Tsujie M, Munemoto Y, Yamaguchi T, Kuroda H, Fukunaga M, Tomita N, Sugihara K. 2021. Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial). *British Journal of Cancer* **124**:1088–1097. DOI: https://doi.org/10.1038/s41416-020-01222-8, PMID: 33414540

Westrich JA, Vermeer DW, Colbert PL, Spanos WC, Pyeon D. 2020. The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. *Molecular Carcinogenesis* **59**:794–806. DOI: https://doi.org/10.1002/mc.23188, PMID: 32212206

Yang M, Li D, Jiang Z, Li C, Ji S, Sun J, Chang Y, Ruan S, Wang Z, Liang R, Dai X, Li B, Zhao H. 2022. TGF-β-Induced FLRT3 Attenuation Is Essential for Cancer-Associated Fibroblast-Mediated Epithelial-Mesenchymal Transition in Colorectal Cancer. *Molecular Cancer Research* **20**:1247–1259. DOI: https://doi.org/10.1158/1541-7786.MCR-21-0924, PMID: 35560224

Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RGW. 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications* **4**:2612. DOI: https://doi.org/10.1038/ncomms3612, PMID: 24113773

Zhang Q, Zhang C, Ma JX, Ren H, Sun Y, Xu JZ. 2019. Circular RNA PIP5K1A promotes colon cancer development through inhibiting miR-1273a. *World Journal of Gastroenterology* **25**:5300–5309. DOI: https://doi.org/10.3748/wjg.v25.i35.5300, PMID: 31558874

[***15***] Zwinsová B, Brychtová V, Hrivňáková M, Zdražilová-Dubská L, Bencsiková B, Šefr R, Nenutil R, Vídeňská P, **Budinská E.** Role of the Microbiome in the Formation and Development of Colorectal Cancer. Klin Onkol. 2019 Summer;32(4):261-269. English. doi: 10.14735/amko2019261. PMID: 31426641. [in Czech]

Vliv mikrobiomu na vznik a vývoj kolorektálního karcinomu

Role of the Microbiome in the Formation and Development of Colorectal Cancer

Zwinsová B., Brychtová V., Hrivňáková M., Zdražilová-Dubská L., Bencsiková B., Šefr R., Nenutil R., Víděňská P., Budinská E.

Regionální centrum aplikované molekulární onkologie, Masarykův onkologický ústav, Brno

Souhrn

Východiska: Kolorektální karcinom je velice heterogenní onemocnění z klinického, histopatologického i molekulárního hlediska. Detailní charakterizace této heterogenity a její vliv na patologii tohoto onemocnění je nezbytným krokem k lepší stratifikaci pacientů a vývoji nových léčebných postupů. V posledních dvou desetiletích se pozornost vědců zaměřovala na studium molekulárních nádorových procesů pro prediktivní, diagnostické a prognostické účely. Avšak i přes veškeré úsilí jsou existující molekulární prediktivní a prognostické testy aplikovatelné pouze pro menší specifické skupiny pacientů a spíše pomáhají při rozhodování o nasazení specializované cílené biologické léčby, než by předpovídaly její úspěšnost. Samotné molekulární profilování není schopné zachytit řadu dalších faktorů významně ovlivňujících růst a agresivitu nádoru. Mezi tyto aspekty patří i mikroprostředí nádoru, jehož nejméně prostudovanou částí je střevní mikrobiom – specifické společenství všech komenzálních, symbiotických a patogenních mikroorganizmů. Střevní mikrobiom hraje klíčovou roli u řady onemocnění, jako je např. Crohnova choroba, diabetes II. typu a obezita a podle nejnovějších studií může dlouhodobá dysbióza střevní mikroflóry ovlivňovat i vznik a další vývoj kolorektálního karcinomu, jeho agresivitu nebo úspěšnost léčby. ***Závěr:*** Tato přehledová studie sumarizuje dosavadní poznatky studia střevního mikrobiomu u kolorektálního karcinomu, vč. různých mechanizmů, jakými střevní mikrobiom ovlivňuje poškození stěny střeva a tím se může podílet na vzniku a progresi kolorektálního karcinomu.

Klíčová slova

kolorektální karcinom – heterogenita – střevní mikrobiom – dysbióza

Summary

Background: The clinical, histopathological, and molecular characteristics of colorectal cancer vary considerably. Factors associated with the heterogeneity of this disease and with understanding the effects of heterogeneity on disease progression and response to therapy are critical for the better stratification of patients and the development of new therapeutic methods. Although studies have focused mainly on tumor molecular profiling, current molecular predictive and prognostic factors are relevant to specific groups of colorectal cancer patients and are mostly used to predict the applicability of targeted biological agents rather than to predict their benefits. Molecular profiling fails to capture aspects important for tumor growth and aggressiveness, including the tumor microenvironment. The gut microbiome, consisting of specific communities of all commensal, symbiotic, and pathogenic microorganisms, has been shown to have a significant impact on the development of many diseases, including Crohn's disease, type II diabetes, and obesity. Recent studies have indicated that long-term dysbiosis of the intestinal microflora can influence the development and progression of colorectal cancer, as well as tumor aggressiveness and response to treatment. ***Conclusion:*** This review article summarizes current knowledge of the gut microbiome in colorectal cancer, including the various mechanisms by which the gut microbiome affects the intestinal wall, thereby contributing to the development and progression of colorectal cancer.

Key words

colorectal cancer – heterogeneity – gut microbiome – dysbiosis

Tato práce byla podpořena Ministerstvem zdravotnictví České republiky (projekt AZV 16-31966A), projektem Ministerstva školství, mládeže a tělovýchovy České republiky – NPU I – LO1413 a Ministerstva zdravotnictví České republiky – RVO (MOÚ, 00209805).

This work was supported by Ministry of Health of the Czech Republic (project AZV 16-31966A), project of Ministry of Education, Youth and Sports of the Czech Republic – NPU I – LO1413 a Ministry of Health of the Czech Republic – RVO (MMCI, 00209805).

Autoři deklarují, že v souvislosti s předmětem studie nemají žádné komerční zájmy.

The authors declare they have no potential conflicts of interest concerning drugs, products, or services used in the study.

Redakční rada potvrzuje, že rukopis práce splnil ICMJE kritéria pro publikace zasílané do biomedicínských časopisů.

The Editorial Board declares that the manuscript met the ICMJE recommendation for biomedical papers.

Mgr. Eva Budinská, Ph.D.
Regionální centrum aplikované molekulární onkologie
Masarykův onkologický ústav
Žlutý kopec 7
656 53 Brno
e-mail: eva.budinska@mou.cz

Obdrženo/Submitted: 15. 4. 2019
Přijato/Accepted: 17. 6. 2019

doi: 10.14735/amko2019261

Úvod

Zhoubný novotvar tlustého střeva a konečníku (colorectal cancer – CRC) je vysoce heterogenní onemocnění, kterému v celosvětovém měřítku patří druhá (ženy) a třetí (muži) příčka v incidenci nádorových onemocnění [1]. V ČR se jedná o jednu z nejčastějších onkologických diagnóz a celosvětově má ČR šestou nejvyšší incidenci CRC [2]. V etiologii tohoto onemocnění hrají významnou roli environmentální vlivy, jako jsou stravovací návyky, konzumace alkoholu a kouření a také rizikové faktory, jako je věk, rodinná anamnéza, chronická zánětlivá onemocnění trávicího traktu a přítomnost polypů tlustého střeva. Přestože incidence CRC aktuálně zaznamenává ve vyspělých zemích mírný pokles především díky zavedení preventivních a screeningových programů [3], jedná se stále o onemocnění s komplikovanou léčbou a vysokou úmrtností. Důvodem je vysoká heterogenita tohoto onemocnění, a to jak interindividuální, tak heterogenita samotného nádoru jediného pacienta. Nádorová heterogenita CRC zahrnuje rozdíly na úrovni fenotypu nádorových buněk (jak morfologické, tak molekulární) a jejich interakci s nádorovým mikroprostředím projevující se přítomností buněk nenádorové tkáně (tzv. stromální reakce), variabilní infiltrací buňkami imunitního systému a – jak naznačují nejnovější výzkumy – také přítomností bakterií v samotném nádoru nebo na jeho povrchu z luminální strany tlustého střeva. Posledních 15 let se vědecký výzkum prognostických a prediktivních markerů CRC (a nádorového onemocnění obecně) zaměřoval zejména na molekulární profilování nádorových buněk a hodnocení faktorů nádorového mikroprostředí. Výsledky tohoto úsilí vedly k zavedení několika molekulárních biomarkerů, které v kombinaci se standardně používanými klinickými proměnnými slouží k individualizaci léčby. Jako příklad můžeme uvést stanovení mikrosatelitové nestability jako pozitivního prognostického markeru při indikaci adjuvantní chemoterapie na bázi 5-fluoruracilu u II. klinického stadia CRC nebo testování mutací genů *KRAS*, *NRAS* a *BRAF* při indikaci léčby metastatického CRC terapeutickými monoklonálními protilátkami cílenými proti receptoru pro epidermální růstový faktor [4]. Dále existuje několik vícegenových prognostických panelů určených pro predikci relapsu u časných stadií CRC jako Oncotype DX Colon Recurrence Score Test [5] a ColoPrint [6], které byly validovány retrospektivně v rámci řady klinických studií [7]. Jejich využití v klinické praxi však není rutinní, protože tyto testy nepredikují benefit adjuvantní chemoterapie. Podobně je na tom prognostický nástroj Immunoscore® [8] založený na kvantifikaci tumor-infiltrujících T buněk v parafinových řezech, který demonstruje význam složení nádorového mikroprostředí pro agresivitu CRC. Ve snaze blíže charakterizovat molekulární procesy CRC se pozornost zaměřila na definici molekulárních podtypů a v roce 2013 bylo publikováno několik klasifikačních systémů vycházejících z transkriptomických profilů CRC [9–12], které byly následně sjednoceny v koordinovaném úsilí za definice konsenzuálních molekulárních podtypů (consensual molecular subtypes – CMS) CRC s rozdílnou genovou expresí a silnou prognostickou hodnotou [13]. V roce 2017 pak byl navržen alternativní systém klasifikace CRC, Current Research Information System [14], založený na transkripčních profilech čisté populace nádorových buněk (bez nádorového stromatu). Oba systémy se vzájemně doplňují, ale neposkytují dostatečná kritéria, která by pomohla s nastavením specifické léčby na míru pacienta. Po 15 letech molekulárních výzkumů v oblasti CRC jsme tedy ještě stále vzdáleni svatému grálu personalizované medicíny. Jedním z pomyslných chybějících střípků mozaiky heterogenity CRC může být právě střevní mikrobiom, který, jak se ukazuje, má významný vliv na vznik a vývoj CRC.

Střevní mikrobiom

Lidské tělo je osídleno řádově triliony symbiotických mikroorganizmů vyskytujících se např. v trávicím traktu, na kůži, sliznici dutiny ústní a slinách, spojivce, dýchacích cestách, urogenitálním systému a dalších [15]. Pojem mikrobiom označuje specifické společenství všech komenzálních, symbiotických i patogenních mikroorganizmů v hostitelském organizmu a jejich genom. Patří sem bakterie, viry, archea a některé eukaryotní organizmy (prvoci, plísně a kvasinky). Každý člověk hostí více mikroorganizmů, než je počet jeho vlastních buněk, a také genetická informace lidského mikrobiomu je několikanásobně větší než genetická informace samotného člověka [16–18]. Z hlediska zdraví je nejvýznamnější mikrobiom tlustého střeva, zejména bakterie, které se podílejí na metabolizmu komplexních složek potravy, které by jinak nebylo možné využít v přeměně na energii, dále na syntéze esenciálních i neesenciálních aminokyselin, enzymů a vitaminů (vitamin K, B12, kyselina listová), můžou neutralizovat potenciálně karcinogenní sloučeniny [19], anebo naopak škodlivé sloučeniny produkovat [20]. Významnými metabolity jsou např. mastné kyseliny s krátkým řetězcem (short chain fatty acids – SCFA), které slouží jako zdroj energie nejen pro kolonocyty, ale i pro ostatní buňky a orgány [21]. Střevní mikrobiom tak ovlivňuje i morfologii střeva – architekturu sliznice, množství a složení hlenu, prokrvenost a proliferaci epiteliálních buněk [22]. V posledních letech se studium střevního mikrobiomu orientuje zejména na jeho interakci s imunitním systémem, která je důležitá obzvláště v období po narození, kdy se významně podílí na vývoji imunity. Pokud nedojde ke správnému vývoji imunitní reakce, dochází k vývoji alergií či autoimunitních chorob. Interakce mezi mikrobiomem a imunitním systémem později napomáhá k rovnováze mezi eliminací patogenů napadajících trávicí trakt a udržením tolerance ke zdravé tkáni střeva [23,24]. Složení střevního mikrobiomu se nejvíce mění od narození přibližně do 3 let věku [25]. V pozdějším věku je složení mikrobiomu stabilní, pokud není narušena jeho rovnováha a nedojde k prudkému snížení diverzity mikrobiomu nebo dominanci bakterií, které nepatří mezi přínosné komenzální zástupce. Tento nerovnovážný stav mikrobiomu nazýváme dysbióza a může být vyvolán např. léčbou antibiotiky, nevhodnou stravou nebo onemocněním. Dysbióza je spojována s vývojem různých chronických a autoimunitních chorob, jako je např. Crohnova choroba, diabetes II. typu, obezita apod. [26].

Střevní mikrobiom je tvořen zejména striktně anaerobními kmeny bakterií, které převažují nad bakteriemi fakultativně anaerobními a aerobními. Protože je velice obtížné tyto bakterie charakterizovat klasickými kultivačními metodami, jejich podrobnější studium umožnil až vývoj molekulárních metod. Vzhledem k velkému množství genetických informací se ve studiu mikrobiomu nejvíce uplatňují sekvenační techniky nové generace, které umožňují číst velké množství genetických informací najednou, a tak identifikovat složení a funkci mikrobiomu. Nejrozšířenější je stále analýza bakteriálního složení pomocí genu kódujícího 16S ribozomální ribonukleové kyseliny (rRNA) se specifickými oblastmi pro jednotlivé bakterie a analýza funkce mikrobiomu pomocí celometagenomového sekvenování [27].

Složení mikrobiomu a abundance jednotlivých bakteriálních taxonů odrážejí fyziologické rozdíly jednotlivých částí gastrointestinálního traktu, a není tedy ve střevním lumen rovnoměrné. Nejvíce bakterií se nachází v tlustém střevě, kde dochází k významnému rozkladu zbytků tráveniny bakteriemi. U zdravého člověka jsou z více než 90 % dominujícími bakteriálními kmeny gram-pozitivní *Firmicutes* a gram-negativní *Bacteroidetes* [28,29], jejichž poměr se významně interindividuálně liší [30]. Mezi další významně zastoupené kmeny patří *Actinobacteria*, *Verrucomicrobia* a *Proteobacteria* [31,32]. Na úrovni řádů jsou dominujícími sacharolytické *Bacteriodales* (kmen *Bacteroidetes*) a *Clostridiales* (kmen *Firmicutes*).

Dopad složení střevního mikrobiomu na zdraví člověka je studován zejména porovnáním složení mikrobiomu u nemocných a zdravých jedinců. Rozsáhlé populační studie jako US National Institute of Health Human Microbiome Project [18] nebo European Metagenomics of the Human Intestinal Tract project [33] ukazují, že ve složení střevního mikrobiomu existuje velká variabilita jak mezi etniky, tak i jedinci, a není proto jednoduché definovat složení „zdravého" mikrobiomu. Snaha o stratifikaci mikrobiomu vyústila v roce 2011 v definování tří základních enterotypů lišících se dominancí jednoho z bakteriálních druhů – *Bacteroides*, *Prevotella* a *Ruminococcus* [34]. Navazující práce ukázaly, že řazení jedinců do enterotypů je velice závislé na použité metodice a není dostatečně ostře vymezené. Nicméně budoucí studie založené na velkých datových souborech s normovanou metodikou mohou tuto problematiku posunout k využití v léčbě a diagnostice nejenom střevních onemocnění [35].

Mikrobiom a kolorektální karcinom

Střevní mikrobiom může, ať už v podobě individuálních bakteriálních zástupců, nebo spolupůsobením mikrobiální komunity, potencovat nebo zmírňovat riziko vzniku CRC. Řada publikací ukazuje, že dysbióza střevní mikroflóry anebo poškození stěny střeva v důsledku interakce mikrobiomu s buňkami imunitního systému hostitele může ovlivňovat rozvoj zánětlivých a nádorových onemocnění střeva, progresi nádorů [36,37] a jejich odpověď na léčbu [38,39]. Bakterie a jejich produkty se mohou podílet na vzniku nebo progresi sporadického CRC řadou různých mechanizmů, jako je indukce prozánětlivých a prokarcinogenních drah v epiteliálních buňkách, produkce genotoxinů a reaktivních forem kyslíku [40,41] nebo metabolická přeměna prokarcinogenních výživových faktorů na karcinogeny [42]. Za významný rizikový faktor vzniku nádorových onemocnění střeva je považována zánětlivá reakce ve střevním epitelu, která může být způsobena dlouhodobou dysbiózou a působením patogenních nebo oportunně (podmíněně) patogenních bakterií [43–45].

Zdravý střevní epitel má řadu obranných mechanizmů pro boj s mikrobiálními narušiteli. Patří sem hlenová vrstva a udržování epiteliální integrity zamezující vniku mikroorganizmů, rychlá výměna epiteliálních buněk, a tedy odstraňování buněk infikovaných, autofagie a vrozená imunitní odpověď. Neméně důležitá je také ochranná funkce komenzálních bakterií, které svojí přítomností na sliznici za normálních okolností zabraňují průniku patogenů nebo jejich produktů. Nicméně mnoho bakteriálních patogenů vyvinulo téměř dokonalé infekční strategie, kterými tyto obranné mechanizmy obcházejí. Během svého působení vylučují různé toxiny a efektory, kterými mohou ovlivňovat hostitelské buněčné funkce a využívat je pro svoje přežití. Zároveň udržují rovnováhu v narušené epiteliální bariéře tak, aby ji mohli kolonizovat dlouhodobě [46]. Vedle patogenních bakterií mohou svými metabolity přispívat ke vzniku CRC také bakterie primárně nepatogenní, jejichž prokarcinogenní potenciál může být umocněn poškozením střevní sliznice následkem zranění nebo infekce [46]. Všechny mechanizmy, kterými střevní mikrobiom jako celek přispívá k udržení zdravého prostředí, nebo naopak rozvoji patologických stavů vč. tumorigeneze v mikrobiálně bohatém a imunologicky komplexním prostředí střeva, ovšem stále nejsou zcela objasněny.

Látky, kterými bakterie negativně působí na buňky epitelu nebo imunitní buňky hostitele, mají cyklomodulační (buněčný cyklus ovlivňující) nebo genotoxický účinek. V důsledku toho může u epiteliálních buněk docházet k poškození deoxyribonukleové kyseliny (DNA), akumulaci mutací a nekontrolované proliferaci. Produkty bakterií se můžou podílet na zastavení dělení imunitních buněk, čímž prakticky umožní sobě i nádorovým buňkám uniknout imunitní odpovědi. Z tohoto pohledu se jeví pro vznik CRC jako nejvýznamnější enterotoxické kmeny *Escherichia coli*, které produkují hned čtyři skupiny cyklomodulinů – cytotoxický nekrotizující faktor (CNF), faktor inhibující buněčný cyklus (Cif), cytoletální distendující toxin (CDT) a kolibaktin, který má také genotoxické účinky. CDT a Cif a kolibaktin inhibují proliferaci, CNF zase proliferaci spouští. CDT kromě enteropatogenních kmenů *E. coli* produkují také *Shigella dysenteriae*, *Campylobacter* spp., *Salmonella typhi*, *Haemophilus ducreyi* nebo *Actinobacillus actinomycetemcomitans* [47]. Enterotoxické kmeny *Bacteroides fragilis* produkují toxin ze skupiny metaloproteáz, který narušuje spojení střevních epiteliálních buněk štěpením transmembránového proteinu E-cadherinu [48–50]. Degradace E-cadherinu má za následek aktivaci β-kateninové signální dráhy v buňkách epitelu, což způsobuje jejich

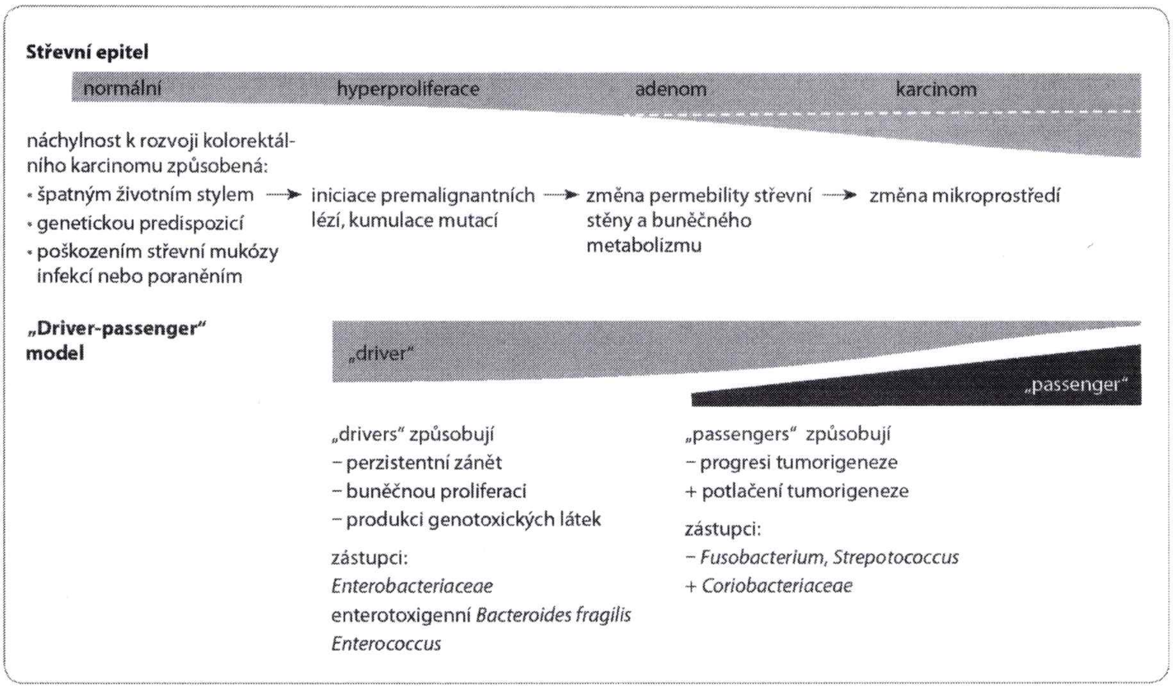

Střevní epitel
normální
hyperproliferace
adenom
karcinom
náchylnost k rozvoji kolorektálního karcinomu způsobená:
• špatným životním stylem
• genetickou predispozicí
• poškozením střevní mukózy infekcí nebo poraněním
iniciace premalignantních lézí, kumulace mutací
změna permebility střevní stěny a buněčného metabolizmu
změna mikroprostředí
„Driver-passenger" model
„driver"
„passenger"
„drivers" způsobují
– perzistentní zánět
– buněčnou proliferaci
– produkci genotoxických látek
zástupci:
Enterobacteriaceae
enterotoxigenní Bacteroides fragilis
Enterococcus
„passengers" způsobují
– progresi tumorigeneze
+ potlačení tumorigeneze
zástupci:
– Fusobacterium, Strepotococcus
+ Coriobacteriaceae

Schéma 1. Bakteriální „driver-passenger" model.

zvýšenou proliferaci. Je zajímavé, že enteropatogenní kmeny *E. coli* jsou také schopny dramaticky snížit expresi klíčových DNA mismatch-repair proteinů MSH2 a MLH1 [51], a mohou dokonce aktivovat senescentní buňky k produkci růstových faktorů, které ovlivňují nádorový růst [52]. Ke komenzálním bakteriím s prokarcinogenním účinkem patří např. *Enterococcus faecalis*, která může poškozovat DNA epiteliálních buněk svou produkcí extracelulárních superoxidů [53,54]. Sulfidogenní bakterie jako *Fusobacterium*, *Desulfovibrio* a *Bilophila wadsworthia* se mohou účastnit procesu nádorového bujení produkcí sirovodíku, jenž je známý svým genotoxickým účinkem [55].

Dalším faktorem, jak se mohou bakterie úspěšněji podílet na progresi procesu tumorigeneze, je tvorba biofilmu – homogenního nebo heterogenního společenstva mikroorganizmů tvořícího vyšší strukturu obklopenou extracelulárními polymerními látkami, které mikroorganizmy v biofilmu vylučují a chrání je tak před nepříznivými podmínkami a imunitou [56,57]. Bakterie schopné adherence na různé povrchy jsou tzv. primární kolonizátoři, ke kterým se později mohou připojit bakterie, které by jinak nebyly samotné adherence schopné – sekundární kolonizátoři [58,59]. Přítomnost biofilmu pozměňuje metabolizmus v nádorové tkáni, produkci regulátorů buněčné proliferace a potenciálně ovlivňuje vývoj a progresi nádoru [60].

Bakteriální model „driver-passenger"

Známý model „adenom–karcinom" vývoje CRC, který byl navržen Fearonem & Vogelsteinem [61], má svou paralelu i v bakteriálním působení na proces tumorigeneze. V modelu adenom–karcinom je pojem „driver–passenger" běžně používán ve spojení s genovými mutacemi, přičemž „driver" jsou mutace řídící, tedy zodpovědné za vznik a vývoj nádoru, a „passenger" jsou mutace vzniklé sekundárně v procesu nádorového vývoje [62]. Tjalsma et al přišli s myšlenkou aplikace tohoto modelu na bakterie a vznik CRC. Hlavní myšlenka tohoto modelu je taková, že vznik a vývoj CRC je iniciován „driver" bakteriemi, které svou aktivitou přímo nebo nepřímo způsobují poškození DNA a které v důsledku mohou přispívat k hromadění mutací charakteristických pro adenomy a karcinomy a k procesu tumorigeneze [63]. Proces nádorového bujení pak mění lokální mikroprostředí, zvyšuje se propustnost střevní stěny a mění se buněčný metabolizmus, což umožní uchycení a pomnožení oportunních „passenger" bakterií. Změna mikroprostředí tedy způsobí, že patogenní bakteriální „drivery" mohou být postupně nahrazeny „passenger" bakteriemi, které buď dále podněcují, nebo naopak pozastaví proces tumorigeneze [63] (schéma 1).

Za bakteriální „drivery" jsou tedy považovány střevní bakterie s prokarcinogenními vlastnostmi. Do této kategorie patří výše uvedené střevní patogenní bakterie. Jako bakteriální „passengers" jsou definované ty střevní bakterie, které relativně málo kolonizují střevo zdravých lidí, ale dokážou se úspěšně množit v mikroprostředí nádoru. Mohou to být oportunní patogeny, komenzály nebo probiotické bakterie a podle toho může dojít buď k progresi, nebo potlačení tumorigeneze [63].

Tento model vznikl na základě řady studií, které porovnávají bakteriální složení ze stolice nebo stěrů střevní tkáně

Tab. 1. Aktuální přehled studií porovnávajících změny ve střevním mikrobiomu ve vzorcích stolice a stěrech nádrové tkáně.

↑ u CRC pacientů	↑ u adenomů	↑ u zdravých kontrol	Typ vzorku	Populace	Reference
Stolice					
Enterococcus Faecalis		*Faecalibacterium prausnitsii*		20 CRC / 17 kontrol	Balamurugan et al, 2008 [95]
Bacteroides, Prevotella			stolice před kolonoskopií	60 CRC / 119 kontrol	Sobhani et al, 2011 [96]
Porphyromonas, Escherichia/Shigella, Enterococcus, Streptococcus, Peptostreptococcus, Bacteroides fragilis		*Bacteroides vulgatus, Bacteroides uniformis, Roseburia,* butyrát produkující bakterie	stolice před operací	46 CRC / 56 kontrol	Wang et al, 2012 [97]
Peptostreptococcus, Mogibacterium, Anaerococcus, Slakia, Paraprevotella, Anaerotruncus, Collinsella, Desulfovibrio, Eubacterium, Porphyromonas			stolice před operací	21 CRC / 22 kontrol	Chen et al, 2012 [98]
Atopobium, Porphyromonas, Fusobacterium		*Ruminococcus*		47 CRC / 94 kontrol	Ahn et al, 2013 [99]
Fusobacterium, Bacteroides		*Faecalibacterium prausnitsii, Roseburia*		19 CRC / 20 kontrol	Wu et al, 2013 [100]
Fusobacterium, Lachnospiraceae, Enterobacteriaceaea, Porphyromonas	*Ruminococcaceae, Clostridium, Pseudomonas, Porphyromonas*	*Blautia, Ruminococcus, Lachnospiraceae, Clostridium, Bacteroides, Clostridiales*	stolice 1–4 týdny po kolonoskopii	30 CRC / 30 adenom / 30 kontrol	Zackular et al, 2014 [64]
Bacteroides, Fusobacterium, Alistipes, Escherichia, Parvimonas, Bilophila		*Ruminococcus, Bifidobacterium, Streptococcus*		46 CRC / 63 kontrol	Feng et al, 2015 [101]
Parvimonas micra, Solobacterium moorei, Peptostreptococcus stomatis				137 CRC / 187 kontrol	Yu et al, 2017 [102]
Stěr z tkáně					
Coriobacteriae, Roseburia, Fusobacterium, Faecalibacterium, butyrát produkující bakterie		*Shigella, Citrobacter, Serratia, Salmonella*	odběr během operace	6 CRC / 6 kontrol	Marchesi et al, 2011 [103]
Bacteroides				22 CRC / 22 kontrol	Sobhani et al, 2011 [96]
Bacteroides, Prevotella, Streptococcus, Fusobacterium, Peptostreptococcus, Morganella, Porphyromonas		*Lactobacillus, Roseburia, Pseudobutyvibrio*	odběr během operace	27 CRC / 27 střevní lumen	Chen et al, 2012 [98]
Fusobacterium, Parvimonas, Gemella, Leptotrichia	*Escherichia coli, Pseudomonas veronii, Lactococcus*		odběr během kolonoskopie	52 CRC / 47 adenom / 61 kontrol	Nakatsu et al, 2015 [104]
Fusobacterium			odběr během operace	55 CRC / 55 přilehlá zdravá tkáň	Viljoen et al, 2015 [71]
Proteobacteria, Bacteroidetes		*Enterococcus, Bacillus, Solibacillus*	odběr během kolonoskopie	31 adenom / 20 kontrol	Lu et al, 2016 [105]

CRC – kolorektální karcinom

u pacientů s CRC, adenomem a zdravých kontrol. V tab. 1 je uveden přehled aktuálních studií a jejich výsledků porovnávajících změny ve stolici a nádorové tkáni. Omezením těchto studií je, že neberou v úvahu možnou rozdílnost v mikrobiálním složení danou podtypy nádorů.

Studie porovnávající bakteriální složení stolice u pacientů s CRC, adenomem a zdravými kontrolami obvykle cílily na využití mikrobiomu stolice jakožto přídatného neinvazivního testu na CRC. V roce 2014 dva vědecké týmy nezávisle na sobě publikovaly studie pojednávající o možnosti využití mikrobiomu jako screeningového nástroje pro diagnostiku začínajícího CRC [64,65]. Oba týmy dospěly k závěru, že kombinací základních rizikových faktorů pro vznik CRC (body mass index, věk a rasa) a mikrobiálního složení stolice lze lépe odlišit skupiny pacientů s CRC, pacientů s adenomem a zdravých lidí.

Mikrobiom, klinické a molekulární proměnné

Další typ studií se zaměřuje na korelaci bakteriálního složení na povrchu nádorové tkáně nebo ve stolici pacientů s CRC s klinickými proměnnými ve snaze pochopit odlišnosti mezi různými podtypy nádorů a podchytit možné biologické pozadí jejich heterogenity.

Lokalizace primárního CRC se ukazuje jako významný klinický faktor, který ovlivňuje charakteristiku nádorů tlustého střeva, a to nejen histopatologickou, ale i molekulární [66]. Je známo, že proximální část střeva je odlišného embryonálního vývoje než jeho distální část [67] a s tím se pojí i rozdíl v cévním zásobování, v expresi antigenů a metabolizmu glukózy, což má za následek odlišné osídlení bakteriemi i u zdravých jedinců [68]. Tento efekt se projevuje také ve významném rozdílu ve složení mikrobiomu mezi pravostrannými a levostrannými nádory tlustého střeva [68]. Proximální část zdravého střeva např. obsahuje větší počet bakterií než část distální [67], ale relativní prevalence dominantních rodů *(Lactococcus, Fusobacterium, Pseudomonas a Flavobacterium)* je v pravé a levé části zdravého střeva podobná. U nádorů je ovšem rod *Fusobacterium* více zastoupen v levé, distální části tlustého střeva [68]. Zde je také více zastoupena bakterie *Escherichia-Shigella* (tyto dva druhy bakterií jsou natolik příbuzné, že je nelze od sebe odlišit), jejíž přítomnost může napomáhat rozvoji karcinogeneze, ale u zdravých jedinců její výskyt převažuje v proximální části. Bakterie rodu *Prevotella*, které jsou spojované se zvýšenou produkcí IL-17 ve střevní sliznici pacientů, dále pak *Selenomonas, Peptostreptoccus* a kmen *Firmicutes* jsou více zastoupeny v nádorech s proximální lokalizací CRC [68]. Rozdíl mezi pravostrannými a levostrannými nádory je i ve výskytu bakteriálního biofilmu. Ve studii Johnsona et al se biofilm vyskytoval v 89 % nádorových tkání a ve všech vzorcích polypů vyskytujících se pravostranně, zatímco u levostranných nádorů se vyskytoval pouze ve 12 % nádorů a v žádném polypu [60].

Další asociace byla nalezena mezi bakteriálním složením střevní mukózy a sporadickými CRC asociovanými s kolitidou, u nichž byl zaznamenán vyšší výskyt bakterií z čeledi *Enterobacteriaceae* a rodu *Sphingomonas* a nižší výskyt *Fusobacterium* a *Ruminococcus* v porovnání se sporadickými CRC neasociovanými s kolitidou [69].

Studií asociujících molekulární proměnné s mikrobiomem zatím není mnoho. Existují publikace, které porovnávají složení mikrobiomu mezi mikrosatelitně stabilními (MSS) CRC a nádory s mikrosatelitní nestabilitou (MSI). MSS nádory vykazují zvýšený výskyt *E. coli* produkující kolibaktin [70], zatímco MSI nádory zvýšený výskyt *Fusobacterium nucleatum* a *E. coli* neprodukující kolibaktin [70–72].

V roce 2017 první a zatím jediná výzkumná skupina publikovala výsledky korelace mikrobiomu s konsenzuálními molekulárními podtypy CRC [73]. Analýza bakteriálního složení proběhla cílenou sekvenací genu pro 16S rRNA z nádoru a zároveň zkoumáním RNA sekvencí samotného nádoru, které nebyly přiřazené k referenčnímu lidskému genomu. Validace pak proběhla pomocí kvantitativní polymerázové řetězové reakce. Tímto způsobem byly identifikovány bakterie, které byly asociovány s konsenzuálními molekulárními podtypy CRC [13]. Zvýšený výskyt bakterií *Fusobacterium hwasooki* a *Porphyromonas gingivalis* je spojován s molekulárním podtypem CMS1 [73]. U tohoto podtypu se také ve větším množství vyskytovaly bakterie *Treponema denticola* a *Tannerrella forsythia*, které se společně s *Porphyromonas gingivalis* podílejí na tvorbě již výše zmíněného biofilmu [74,75]. Druhý molekulární podtyp (CMS2 Canonical) obsahoval zvýšený výskyt bakterií *Selenomonas* a *Prevotella*. *Bacillus coagulants* je asociován s třetím molekulárním podtypem (CMS3 Metabolic). Tato studie je ovšem limitovaná malým množstvím vzorků (n = 34), na kterých byla analýza provedena, a faktem, že ani jeden vzorek nebyl zařazen do čtvrtého konsenzuálního podtypu (CMS4 Mesenchymal), který je navíc prognosticky nejvýznamnější [73].

Funkční studie bakterie *Fusobacterium nucleatum*

Často pozorované asociace zvýšeného výskytu bakterií *Fusobacterium nucleatum* ve stolici i nádorové tkáni u pacientů s CRC vedl k řadě následných detailnějších studií, a tak se z bakterie *F. nucleatum* stal zřejmě nejlépe popsaný druh tohoto bakteriálního rodu z pohledu CRC. Jedná se o bakterie podporující zánět, které se abundantně vyskytují ve střevním mikrobiomu pacientů se zánětlivým střevním onemocněním [76]. Přítomnost *F. nucleatum* podporuje expresi celé řady zánětlivých cytokinů, např. IL-6, IL-8, IL-10, IL-18, TNF-α [72–74]. *F. nucleatum* usnadňuje díky ojedinělým dvěma typům adherence spojení primárních a sekundárních kolonizátorů a podílí se tak výrazným způsobem na vzniku biofilmu [58]. Schopnost *F. nucleatum* zachytit se na střevní sliznici nebo do ní invadovat je závislá na formě adhezivního proteinu FadA, který umožňuje jejich přilnutí k epitelovým buňkám střevní sliznice. Současně dochází k sérii událostí, které vyústí ve stimulaci proliferace CRC buněk [77]. Kromě uvedeného dochází k snížení sekrece mucinů pohárkovými buňkami střeva jakožto hlavní složky slizniční ochrany střeva [78–81].

Vyšší výskyt *F. nucleatum* v nádorové tkáni koreluje s nižší infiltrací nádorových lézí T lymfocyty, vyšším stadiem onemocnění a horší prognózou z důvodu vyšší agresivity těchto nádorů [72,82]. Zdá se, že jedním z důvodů, proč dochází k potlačování imunitní odpovědi organizmu v místě nádorové léze napadené *F. nucleatum*, je skutečnost, že metabolity této bakterie jsou krátké peptidy a SCFA, které slouží jako chemoatraktant pro myeloidní supresorové buňky [83]. Je zajímavé, že výskyt *F. nucleatum* je spojován s molekulárním podtypem CRC, který je charakteristický vysokou MSI, mutací BRAF a s pravostrannou anatomickou lokalizací [72,84], dále pak s hypermetylací MLH1 a vysokou metylací CpG (CIMP-H) [72].

V roce 2017 provedli Bullman et al [85] pokus na myších, jimž byl zaveden štěp pocházející z primárního CRC získaného od pacientů. Tyto myši byly následně léčeny antibiotiky s cílem snížit množství *F. nucleatum*. Následně bylo pozorováno zpomalení proliferace nádorových buněk a růstu nádoru [85].

Nedávno byl identifikován nový druh bakterie *Fusobacterium*, *Fusobacterium hwasookii*, která byla ve studiích založených na cílené sekvenaci 16S rRNA genu, která má menší přesnost určení do druhů, vždy určována jako *F. nucleatum* [86]. Na základě podobnosti v sekvencích *F. hwasookii* a *F. nucleatum* a výskytu vysoce konzervovaného genu *FadA* u *F. hwasookii* se zdá, že *F. hwasookii* má podobný význam v karcinogenezi jako *F. nucleatum*.

Mikrobiom a léčba kolorektálního karcinomu

Rovnováha mezi bakteriemi je důležitá pro ochranu sliznice střeva a změny vyvolané chemoterapií mohou zvyšovat riziko výskytu infekce. Chemoterapie i radiační léčba ovlivňuje složení střevního mikrobiomu a působí toxicky na střevní sliznici, čímž umožňuje prostup bakterií do vrstvy epitelu. Zwielehner et al [87] ukázali, že množství bakterií klesá u pacientů ihned po aplikování chemoterapie a k obnovení dochází po 5–9 dnech. Chemoterapie významně ovlivnila výskyt komenzálních bakterií z rodu *Bacteroides* a *Bifidobacteria* a potenciálních patogenů z rodu *Clostridium* skupina IV [87]. Přemnožení potenciálních patogenů, jako je např. *Clostridium difficile*, může vést u pacientů s nádorovým onemocněním k závažným komplikacím [88,89].

Protinádorová léčba má vliv na složení mikrobiomu a i složení mikrobiomu může mít vliv na účinnost léčby. Studie na myších modelech odhalují, že střevní mikrobiom má vliv na úspěšnost chemoterapie tím, že ovlivňuje diferenciaci a funkci myeloidních buněk v mikroprostředí nádoru [38]. V této studii Goldszmid et al zkoumali vliv mikrobiomu na léčbu platinovými deriváty, jako je např. oxaliplatina nebo cisplatina. Tyto sloučeniny se vážou na DNA buněk, narušují její strukturu a tím inhibují syntézu a proliferaci proteinů a indukují apoptózu [90]. Oxaliplatina navíc indukuje imunogenní buněčnou smrt, která pak aktivuje protinádorovou imunitu zprostředkovanou T lymfocyty [91]. Léčebný efekt cisplatiny a oxaliplatiny byl mnohem nižší u myší léčených antibiotiky a bezmikrobních myší. U myší léčených antibiotiky se vytvořily sloučeniny platiny a DNA nádorových buněk na stejné úrovni jako u kontrolních myší, avšak již 48 hod po léčbě došlo k významnému poklesu cytotoxicity. Antibiotická léčba myší potlačila veškerou modifikaci genové exprese indukovanou oxaliplatinou [38].

Závěr

Interakce hostitel-mikrobiom se zvláštním důrazem na roli střevního mikrobiomu nejen v imunitní homeostáze a autoimunitních onemocněních, ale i ve spojení se vznikem nádorů bude velmi diskutovaným tématem následujících let. Poslední desetiletí ve výzkumu CRC bylo věnováno především studiu prognostických a prediktivních markerů na základě genové exprese. Nicméně tyto poznatky oproti standardním klinickým faktorům v porovnání s vynaloženým úsilím na jejich získání jen nepatrně zasáhly do klinické praxe. Vezmeme-li v potaz nejnovější znalosti o vlivu mikrobiomu na vznik a vývoj CRC, dá se předpokládat, že studium mikrobiomu může vést k lepšímu porozumění heterogenity CRC a přiblíží nás více k efektivní diagnostice a léčbě. Navíc se v posledních pár letech ukazuje, že nejen bakterie v nádoru, ale i bakterie na jeho povrchu tvořící bakteriální biofilm ovlivňují vývoj CRC. Výsledky studií je však třeba interpretovat velmi opatrně, neboť výsledky jsou závislé na typu odběru, uchování vzorků a izolace bakteriální DNA ze vzorku [92–94].

Z klinického hlediska existuje několik rozdílných terapeutických přístupů, které by potenciálně mohly být použity jako prevence vzniku nádorových onemocnění, jako podpůrná terapie při léčbě nádorových onemocnění anebo ke zvýšení odpovědi na léčbu. Mezi tyto přístupy patří užívání probiotik a prebiotik, změna stravování nebo mikrobiální transplantace. Poslední z výše zmiňovaných ukazuje velmi slibné výsledky při léčbě infekce bakterií *Clostridium difficile* u lidí a je navrhovaná jako léčebný postup v případě idiopatických střevních zánětů a metabolických poruch [39].

Ačkoliv pozorujeme velký pokrok v oblasti výzkumu mikrobiomu, což dokazuje i rostoucí počet publikací na toto téma, stále existuje mnoho nezodpovězených otázek souvisejících s vlivem mikrobiomu na rozvoj CRC. V tomto směru zatím chybějí studie, které se na mikrobiom dívají ze širší perspektivy, vč. nádorového mikroprostředí a jeho klinické i molekulární heterogenity, a poskytly by tak komplexní náhled na vlivy vedoucí ke vzniku a vývoji CRC.

Literatura

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2018. CA Cancer J Clin 2018; 68(1): 7–30. doi: 10.3322/caac.21442.

2. Novotvary 2015. Ústav zdravotnických informací a statistik ČR. [online]. Dostupné z: https://www.uzis.cz/publikace/novotvary-2015.

3. Arnold M, Sierra MS, Laversanne M et al. Global patterns and trends in colorectal cancer incidence and mortality. Gut 2017; 66(4): 683–691. doi: 10.1136/gutjnl-2015-310912.

4. Tan C, Du X. KRAS mutation testing in metastatic colorectal cancer. World J Gastroenterol 2012; 18(37): 5171–5180. doi: 10.3748/wjg.v18.i37.5171.

5. Renfro LA, Zhang N, Lopatin M et al. Prospective evaluation of a 12-gene assay on patient treatment decisions and physician confidence in mismatch repair proficient stage IIA colon cancer. Clin Colorectal Cancer 2017; 16(1): 23–30. doi: 10.1016/j.clcc.2016.07.016.

6. Kopetz S, Tabernero J, Rosenberg R et al. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. Oncologist 2015; 20(2): 127–133. doi: 10.1634/theoncologist.2014-0325.

7. Sharif S, O'Connell MJ. Gene signatures in stage II colon cancer: a clinical review. Curr Colorectal Cancer Rep 2012; 8(3): 225–231. doi: 10.1007/s11888-012-0132-7.

8. Galon J, Pagès F, Marincola FM et al. Cancer classification using the immunoscore: a worldwide task force. J Transl Med 2012; 10: 205. doi: 10.1186/1479-5876-10-205.

9. Budinska E, Popovici V, Tejpar S et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 2013; 231(1): 63–76. doi: 10.1002/path.4212.

10. Dienstmann R, Guinney J, Delorenzi M et al. Colorectal cancer subtyping consortium (CRCSC) identification of a consensus of molecular subtypes. J Clin Oncol 2014; 32 (Suppl 15): 3511–3511.

11. Marisa L, Reyniès A de, Duval A et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013; 10(5): e1001453. doi: 10.1371/journal.pmed.1001453.

12. Sadanandam A, Lyssiotis CA, Homicsko K et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 2013; 19(5): 619–625. doi: 10.1038/nm.3175.

13. Guinney J, Dienstmann R, Wang X et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015; 21(11): 1350–1356. doi: 10.1038/nm.3967.

14. Isella C, Brundu F, Bellomo SE et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun 2017; 8: 15107. doi: 10.1038/ncomms15107.

15. Li K, Bihan M, Yooseph S et al. Analyses of the microbial diversity across the human microbiome. PloS One 2012; 7(6): e32118. doi: 10.1371/journal.pone.0032118.

16. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 2016; 14(8): e1002533. doi: 10.1371/journal.pbio.1002533.

17. Sender R, Fuchs S, Milo R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell 2016; 164(3): 337–340. doi: 10.1016/j.cell.2016.01.013.

18. Turnbaugh PJ, Ley RE, Hamady M et al. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 2007; 449(7164): 804–810. doi: 10.1038/nature06244.

19. Jandhyala SM, Talukdar R, Subramanyam C et al. Role of the normal gut microbiota. World J Gastroenterol 2015; 21(29): 8787–8803. doi: 10.3748/wjg.v21.i29.8787.

20. Koeth RA, Wang Z, Levison BS et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 2013; 19(5): 576–585. doi: 10.1038/nm.3145.

21. LeBlanc JG, Chain F, Martín R et al. Beneficial effects on host energy metabolism of short-chain fatty acids and vitamins produced by commensal and probiotic bacteria. Microb Cell Factories 2017; 16(1): 79. doi: 10.1186/s12934-017-0691-z.

22. Sommer F, Bäckhed F. The gut microbiota – masters of host development and physiology. Nat Rev Microbiol 2013; 11(4): 227–238. doi: 10.1038/nrmicro2974.

23. Ho JT, Chan GC, Li JC. Systemic effects of gut microbiota and its relationship with disease and modulation. BMC Immunol 2015; 16: 21. doi: 10.1186/s12865-015-0083-2.

24. Wu HJ, Wu E. The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes 2012; 3(1): 4–14. doi: 10.4161/gmic.19320.

25. Hill CJ, Lynch DB, Murphy K et al. Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET cohort. Microbiome 2017; 5(1): 4. doi: 10.1186/s40168-016-0213-y.

26. Jurjus A, Eid A, Al Kattar S et al. Inflammatory bowel disease, colorectal cancer and type 2 diabetes mellitus: the links. BBA Clin 2015; 5: 16–24. doi: 10.1016/j.bbacli.2015.11.002.

27. Di Bella JM, Bao Y, Gloor GB et al. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods 2013; 95(3): 401–414. doi: 10.1016/j.mimet.2013.08.011.

28. Eckburg PB. Diversity of the human intestinal microbial flora. Science 2005; 308(5728): 1635–1638. doi: 10.1126/science.1110591.

29. Galperin MY. Genome diversity of spore-forming firmicutes. Microbiol Spectr 2013; 1(2). doi: 10.1128/microbiolspectrum.TBS-0015-2012.

30. Qin J, Li R, Raes J et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464(7285): 59–65. doi: 10.1038/nature08821.

31. Kang M, Martin A. Microbiome and colorectal cancer: unraveling host-microbiota interactions in colitis-associated colorectal cancer development. Semin Immunol 2017; 32: 3–13. doi: 10.1016/j.smim.2017.04.003.

32. Serban DE. Gastrointestinal cancers: influence of gut microbiota, probiotics and prebiotics. Cancer Lett 2014; 345(2): 258–270. doi: 10.1016/j.canlet.2013.08.013.

33. Ehrlich SD. MetaHIT: The European Union project on metagenomics of the human intestinal tract. In: Nelson KE (ed). Metagenomics of the human body. New York: Springer New York 2011: 307–316.

34. Arumugam M, Raes J, Pelletier E et al. Enterotypes of the human gut microbiome. Nature 2011; 473(7346): 174–180. doi: 10.1038/nature09944.

35. Costea PI, Hildebrand F, Arumugam M et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 2018; 3(1): 8–16. doi: 10.1038/s41564-017-0072-8.

36. Chen J, Pitmon E, Wang K. Microbiome, inflammation and colorectal cancer. Semin Immunol 2017; 32: 43–53. doi: 10.1016/j.smim.2017.09.006.

37. Dulal S, Keku TO. Gut microbiome and colorectal adenomas. Cancer J 2014; 20(3): 225–231. doi: 10.1097/PPO.0000000000000050.

38. Goldszmid RS, Dzutsev A, Viaud S et al. Microbiota modulation of myeloid cells in cancer therapy. Cancer Immunol Res 2015; 3(2): 103–109. doi: 10.1158/2326-6066.CIR-14-0225.

39. Dzutsev A, Goldszmid RS, Viaud S et al. The role of the microbiota in inflammation, carcinogenesis, and cancer therapy. Eur J Immunol 2015; 45(1): 17–3. doi: 10.1002/eji.201444972.

40. Lax AJ. Opinion: Bacterial toxins and cancer – a case to answer? Nat Rev Microbiol 2005; 3(4): 343–349. doi: 10.1038/nrmicro1130.

41. Song M, Garrett WS, Chan AT. Nutrients, foods, and colorectal cancer prevention. Gastroenterology 2015; 148(6): 1244–1260. doi: 10.1053/j.gastro.2014.12.035.

42. Bonnet M, Buc E, Sauvanet P et al. Colonization of the human gut by E. coli and colorectal cancer risk. Clin Cancer Res 2014; 20(4): 859–867. doi: 10.1158/1078-0432.CCR-13-1343.

43. Arthur JC, Perez-Chanona E, Mühlbauer M et al. Intestinal inflammation targets cance – inducing activity of the microbiota. Science 2012; 338(6103): 120–123. doi: 10.1126/science.1224820.

44. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell 2010; 140(6): 883–899. doi: 10.1016/j.cell.2010.01.025.

45. Sun J, Kato I. Gut microbiota, inflammation and colorectal cancer. Genes Dis 2016; 3(2): 130–143. doi: 10.1016/j.gendis.2016.03.004.

46. Kim M, Ashida H, Ogawa M et al. Bacterial interactions with the host epithelium. Cell Host Microbe 2010; 8(1): 20–35. doi: 10.1016/j.chom.2010.06.006.

47. De Rycke J, Oswald E. Cytolethal distending toxin (CDT): a bacterial weapon to control host cell proliferation? FEMS Microbiol Lett 2001; 203(2): 141–148. doi: 10.1111/j.1574-6968.2001.tb10832.x.

48. Wu S, Lim KC, Huang J et al. Bacteroides fragilis enterotoxin cleaves the zonula adherens protein, E-cadherin. Proc Natl Acad Sci USA 1998; 95(25): 14979–14984. doi: 10.1073/pnas.95.25.14979.

49. Wu S, Morin PJ, Maouyo D et al. Bacteroides fragilis enterotoxin induces c-Myc expression and cellular proliferation. Gastroenterology 2003; 124(2): 392–400. doi: 10.1053/gast.2003.50047.

50. Sears CL, Geis AL, Housseau F. Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. J Clin Invest 2014; 124(10): 4166–4172. doi: 10.1172/JCI72334.

51. Maddocks ODK, Short AJ, Donnenberg MS et al. Attaching and effacing Escherichia coli downregulate DNA mismatch repair protein in vitro and are associated with colorectal adenocarcinomas in humans. PLoS One 2009; 4(5): e5517. doi: 10.1371/journal.pone.0005517.

52. Cougnoux A, Dalmasso G, Martinez R et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. Gut 2014; 63(12): 1932–1942. doi: 10.1136/gutjnl-2013-305257.

53. Wang X, Huycke MM. Extracellular superoxide production by Enterococcus faecalis promotes chromosomal instability in mammalian cells. Gastroenterology 2007; 132(2): 551–561. doi: 10.1053/j.gastro.2006.11.040.

54. Wang X, Allen TD, May RJ et al. Enterococcus faecalis induces aneuploidy and tetraploidy in colonic epithelial cells through a bystander effect. Cancer Res 2008; 68(23): 9909–9917. doi: 10.1158/0008-5472.CAN-08-1551.

55. Attene-Ramos MS, Wagner ED, Plewa MJ et al. Evidence that hydrogen sulfide is a genotoxic agent. Mol Cancer Res 2006; 4(1): 9–14. doi: 10.1158/1541-7786.MCR-05-0126.

56. Dejea CM, Fathi P, Craig JM et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science 2018; 359(6375): 592–597. doi: 10.1126/science.aah3648.

57. Dickson I. Colorectal cancer: Bacterial biofilms and toxins prompt a perfect storm for colon cancer. Nat Rev Gastroenterol Hepatol 2018; 15(3): 129. doi: 10.1038/nrgastro.2018.16.

58. Kaplan CW, Lux R, Haake SK et al. The Fusobacterium nucleatum outer membrane protein RadD is an arginine-inhibitable adhesin required for inter-species adherence and the structured architecture of multispecies biofilm. Mol Microbiol 2009; 71(1): 35–47. doi: 10.1111/j.1365-2958.2008.06503.x.

59. Li S, Konstantinov SR, Smits R et al. Bacterial biofilms in colorectal cancer initiation and progression. Trends Mol Med 2017; 23(1): 18–30. doi: 10.1016/j.molmed.2016.11.004.

60. Johnson CH, Dejea CM, Edler D et al. Metabolism links bacterial biofilms and colon carcinogenesis. Cell Metab 2015; 21(6): 891–897. doi: 10.1016/j.cmet.2015.04.011.

61. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990; 61(5): 759–767. doi: 10.1016/0092-8674(90)90186-i.

62. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet 1993; 9(4): 138–141.

63. Tjalsma H, Boleij A, Marchesi JR et al. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol 2012; 10(8): 575–582. doi: 10.1038/nrmicro2819.

64. Zackular JP, Rogers MA, Ruffin MT et al. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res 2014; 7(11): 1112–1121. doi: 10.1158/1940-6207.CAPR-14-0129.

65. Zeller G, Tap J, Voigt AY et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014; 10: 766. doi: 10.15252/msb.20145645.

66. Richman S, Adlard J. Left and right sided large bowel cancer. BMJ 2002; 324(7343): 931–932. doi: 10.1136/bmj.324.7343.931.

67. Hagland HR, Søreide K. Cellular metabolism in colorectal carcinogenesis: influence of lifestyle, gut microbiome and metabolic pathways. Cancer Lett 2015; 356(2PtA): 273–280. doi: 10.1016/j.canlet.2014.02.026.

68. Gao Z, Guo B, Gao R et al. Microbiota disbiosis is associated with colorectal cancer. Front Microbiol 2015; 6: 20. doi: 10.3389/fmicb.2015.00020.

69. Richard ML, Liguori G, Lamas B et al. Mucosa-associated microbiota dysbiosis in colitis associated cancer. Gut Microbes 2018; 9(2): 131–142. doi: 10.1080/19490976.2017.1379637.

70. Gagnière J, Bonnin V, Jarrousse AS et al. Interactions between microsatellite instability and human gut colonization by Escherichia coli in colorectal cancer. Clin Sci (Lond) 2017; 131(6): 471–485. doi: 10.1042/CS20160876.

71. Viljoen KS, Dakshinamurthy A, Goldberg P et al. Quantitative profiling of colorectal cancer-associated bacteria reveals associations between fusobacterium spp., enterotoxigenic bacteroides fragilis (ETBF) and clinicopathological features of colorectal cancer. PLoS One 2015; 10(3): e0119462. doi: 10.1371/journal.pone.0119462.

72. Mima K, Nishihara R, Qian ZR et al. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. Gut 2016; 65(12): 1973–1980. doi: 10.1136/gutjnl-2015-310101.

73. Purcell RV, Visnovska M, Biggs PJ et al. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci Rep 2017; 7(1): 11590. doi: 10.1038/s41598-017-11237-6.

74. Grenier D. Demonstration of a bimodal coaggregation reaction between Porphyromonas gingivalis and Treponema denticola. Oral Microbiol Immunol 1992; 7(5): 280–284.

75. Meuric V, Martin B, Guyodo H et al. Treponema denticola improves adhesive capacities of Porphyromonas gingivalis. Mol Oral Microbiol 2013; 28(1): 40–53. doi: 10.1111/omi.12004.

76. Han YW. Fusobacterium nucleatum: a commensal-turned pathogen. Curr Opin Microbiol 2015; 0: 141–147. doi: 10.1016/j.mib.2014.11.013.

77. Rubinstein MR, Wang X, Liu W et al. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/b-catenin signaling via its FadA adhesin. Cell Host Microbe 2013; 14(2): 195–206. doi: 10.1016/j.chom.2013.07.012.

78. Allen-Vercoe E, Strauss J, Chadee K. Fusobacterium nucleatum: an emerging gut pathogen? Gut Microbes 2011; 2(5): 294–298. doi: 10.4161/gmic.2.5.18603.

79. Krisanaprakornkit S, Kimball JR, Weinberg A et al. Inducible expression of human beta-defensin 2 by Fusobacterium nucleatum in oral epithelial cells: multiple signaling pathways and role of commensal bacteria in innate immunity and the epithelial barrier. Infect Immun 2000; 68(5): 2907–2915. doi: 10.1128/iai.68.5.2907-2915.2000.

80. Moore RA, Warren RL, Freeman JD et al. The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS One 2011; 6(5): e19838. doi: 10.1371/journal.pone.0019838.

81. Strauss J, Kaplan GG, Beck PL et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. Inflamm Bowel Dis 2011; 17(9): 1971–1978. doi: 10.1002/ibd.21606.

82. Mima K, Sukawa Y, Nishihara R et al. Fusobacterium nucleatum and T cells in colorectal carcinoma. JAMA Oncol 2015; 1(5): 653–661. doi: 10.1001/jamaoncol.2015.1377.

83. Anand S, Kaur H, Mande SS. Comparative in silico analysis of butyrate production pathways in gut commensals and pathogens. Front Microbiol 2016; 7: 1945. doi: 10.3389/fmicb.2016.01945.

84. Xiao Y, Freeman GJ. The microsatellite instable subset of colorectal cancer is a particularly good candidate for checkpoint blockade immunotherapy. Cancer Discov 2015; 5(1): 16–18. doi: 10.1158/2159-8290.CD-14-1397.

85. Bullman S, Pedamallu CS, Sicinska E et al. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science 2017; 358(6369): 1443–1448. doi: 10.1126/science.aal5240.

86. Cho E, Park SN, Lim YK et al. Fusobacterium hwasookii sp. nov., isolated from a human periodontitis lesion. Curr Microbiol 2015; 70(2): 169–175. doi: 10.1007/s00284-014-0692-7.

87. Zwielehner J, Lassl C, Hippe B et al. Changes in human fecal microbiota due to chemotherapy analyzed by TaqMan-PCR, 454 sequencing and PCR-DGGE fingerprinting. PloS One 2011; 6(12): e28654. doi: 10.1371/journal.pone.0028654.

88. Guarner F, Malagelada JR. Gut flora in health and disease. The Lancet 2003; 361(9356): 512–519. doi: 10.1016/S0140-6736(03)12489-0.

89. van Vliet MJ, Tissing WJ, Dun CA et al. Chemotherapy treatment in pediatric patients with acute myeloid leukemia receiving antimicrobial prophylaxis leads to a relative increase of colonization with potentially pathogenic bacteria in the gut. Clin Infect Dis 2009; 49(2): 262–270. doi: 10.1086/599346.

90. Siddik ZH. Cisplatin: mode of cytotoxic action and molecular basis of resistance. Oncogene 2003; 22(47): 7265–7279. doi: 10.1038/sj.onc.1206933.

91. Kroemer G, Galluzzi L, Kepp O et al. Immunogenic cell death in cancer therapy. Annu Rev Immunol 2013; 31: 51–72. doi: 10.1146/annurev-immunol-032712-100008.

92. Tedjo DI, Jonkers DM, Savelkoul PH et al. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. PloS One 2015; 10(5): e0126685. doi: 10.1371/journal.pone.0126685.

93. Mathay C, Hamot G, Henry E et al. Method optimization for fecal sample collection and fecal DNA extraction. Biopreservation Biobanking 2015; 13(2): 79–93. doi: 10.1089/bio.2014.0031.

94. Panek M, Čipčić Paljetak H, Barešić A et al. Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. Sci Rep 2018; 8(1): 5143. doi: 10.1038/s41598-018-23296-4.

95. Balamurugan R, Rajendiran E, George S et al. Real-time polymerase chain reaction quantification of specific butyrate-producing bacteria, Desulfovibrio and Enterococcus faecalis in the feces of patients with colorectal cancer. J Gastroenterol Hepatol 2008; 23(8 Pt 1): 1298–1303. doi: 10.1111/j.1440-1746.2008.05490.x.

96. Sobhani I, Tap J, Roudot-Thoraval F et al. Microbial dysbiosis in colorectal cancer (CRC) patients. PloS One 2011; 6(1): e16393. doi: 10.1371/journal.pone.0016393.

97. Wang T, Cai G, Qiu Y et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J 2012; 6(2): 320–329. doi: 10.1038/ismej.2011.109.

98. Chen W, Liu F, Ling Z et al. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PloS One 2012; 7(6): e39743. doi: 10.1371/journal.pone.0039743.

99. Ahn J, Sinha R, Pei Z et al. Human gut microbiome and risk for colorectal cancer. J Natl Cancer Inst 2013; 105(24): 1907–1911. doi: 10.1093/jnci/djt300.

100. Wu N, Yang X, Zhang R et al. Dysbiosis signature of fecal microbiota in colorectal cancer patients. Microb Ecol 2013; 66(2): 462–470. doi: 10.1007/s00248-013-0245-9.

101. Feng Q, Liang S, Jia H et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 2015; 6: 6528. doi: 10.1038/ncomms7528.

102. Yu J, Feng Q, Wong SH et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 2017; 66(1): 70–78. doi: 10.1136/gutjnl-2015-309800.

103. Marchesi JR, Dutilh BE, Hall N et al. Towards the human colorectal cancer microbiome. PloS One 2011; 6(5): e20447. doi: 10.1371/journal.pone.0020447.

104. Nakatsu G, Li X, Zhou H et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat Commun 2015; 6: 8727. doi: 10.1038/ncomms9727.

105. Lu Y, Chen J, Zheng J et al. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. Sci Rep 2016; 6: 26337. doi: 10.1038/srep26337.

[***16***] Videnska P, Smerkova K, Zwinsova B, Popovici V, Micenkova L, Sedlar K, **Budinska E**. Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform. Sci Rep. 2019 Sep 25;9(1):13837. doi: 10.1038/s41598-019-49520-3. PMID: 31554833; PMCID: PMC6761292.

SCIENTIFIC REPORTS
natureresearch

OPEN

Received: 14 November 2018
Accepted: 22 August 2019
Published online: 25 September 2019

Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform

Petra Videnska[1], Kristyna Smerkova[1], Barbora Zwinsova[1], Vlad Popovici[1], Lenka Micenkova[1], Karel Sedlar[2] & Eva Budinska[1]

Many studies correlate changes in human gut microbiome with the onset of various diseases, mostly by 16S rRNA gene sequencing. Setting up the optimal sampling and DNA isolation procedures is crucial for robustness and reproducibility of the results. We performed a systematic comparison of several sampling and DNA isolation kits, quantified their effect on bacterial gDNA quality and the bacterial composition estimates at all taxonomic levels. Sixteen volunteers tested three sampling kits. All samples were consequently processed by two DNA isolation kits. We found that the choice of both stool sampling and DNA isolation kits have an effect on bacterial composition with respect to Gram-positivity, however the isolation kit had a stronger effect than the sampling kit. The proportion of bacteria affected by isolation and sampling kits was larger at higher taxa levels compared to lower taxa levels. The PowerLyzer PowerSoil DNA Isolation Kit outperformed the QIAamp DNA Stool Mini Kit mainly due to better lysis of Gram-positive bacteria while keeping the values of all the other assessed parameters within a reasonable range. The presented effects need to be taken into account when comparing results across multiple studies or computing ratios between Gram-positive and Gram-negative bacteria.

The gut microbiome plays a key role in shaping human health and has been the subject of an increasing number of studies in the context of disease development, diagnostics and treatment. Important progress has been made especially in investigating uncultured bacteria, which constitute the main part of the gut microbiome and were previously difficult to characterize with standard techniques such as cloning, Sanger sequencing or Denaturing Gradient Gel Electrophoresis (DGGE)[1,2]. Next generation sequencing (NGS) provides new and more detailed means to study the human microbiome and helps uncovering its impact on the human immune system development[3–5], or on the development of chronic diseases[6,7]. However, human microbiome is very dynamic and can change rapidly in response to many factors such as diet, antibiotic use, lifestyle or environment[8–16]. Many diseases were associated with a phenomenon called dysbiosis – microbial imbalance. Unfortunately, due to the huge microbiome variability it is very difficult to define a normality baseline for an individual. To extract disease-relevant information and generate new or confirm existing biological hypotheses, large cohort microbiome studies are needed. These studies face multiple challenges with the microbiome sampling. First, successful compliance of participants with the established protocol demands both motivation and an easy sampling procedure. Especially, sampling of the stool at home can induce a "yuck effect" and positive education and uncomplicated sampling workflow can significantly decrease the number of study drop-outs[17,18].

Another major problem is the large variability of methodological approaches employed by different microbiome studies. The final composition of bacteria as assessed by sequencing the 16S rRNA gene is influenced by many factors: sampling method[19–22], sample storage conditions[20,22–29], DNA extraction[8,21,22,26,30–39], primers

[1]RECETOX, Faculty of Science, Masaryk University, Kamenice 5, 625 00, Brno, Czech Republic. [2]Department of Biomedical Engineering, Brno University of Technology, Technicka 12, Brno, Czech Republic. Correspondence and requests for materials should be addressed to E.B. (email: budinska@recetox.muni.cz)

targeting different parts of the 16S rRNA gene[40,41] and data analysis[42]. All of these factors may lead to the misinterpretation of changes in the microbiome and thus hamper direct comparisons of results between individual studies[43–45]. These technical problems, along with an as yet unknown gut microbiome diversity in the healthy population, lead to challenges in the implementation of metagenomics into cohort studies and, in consequence, delay the translation of the knowledge to clinical practice.

Most studies focused on the technical factors influencing the assessment of bacterial composition often provide only a description of the observed differences on a limited number of samples, while the comparison of the effect sizes of these factors, or combination thereof remains unexplored. The effect of sampling was previously described with respect to storage conditions (such as temperatures[20,23,26,28,29], periods at room temperature[20,24] or a presence and type of stabilizer[19,21,22,27,28]). None of these studies reported on the volunteers' compliance or the differences in preprocessing steps specific to different sampling kits. Multiple studies describe the effect of stool homogenization prior DNA extraction[25,46], but they only report its overall effect on the interindividual variation, without quantifying this effect at different bacterial taxon levels.

The DNA extraction method was highlighted as a critical factor influencing the observed bacterial composition[39,47]. Commercially available extraction kits use different lysis procedures such as enzymatic, chemical or mechanical bacterial cell disruption methods. Generally, the combination of enzymatic and mechanical disruption is recommended as more effective in the lysis of Gram-positive bacteria[8,22,26,34,35,37,39]. However, these DNA extraction comparison studies are limited to a rather small number of individuals (from 2 to 9) and none of them compared the kits in terms of DNA yield and quality, presence of PCR inhibitors, the human to bacterial DNA ratio, the efficiency of Gram-positive bacteria cell wall lysis and the observed bacterial composition at different taxa levels all at once.

The aim of our study was therefore to perform systematic assessment of effect of sampling and DNA isolation kits and their combinations on a full range of parameters of bacterial DNA quality, bacterial diversity and composition, with respect to user acceptance.

Results

We analyzed stool samples from sixteen volunteers. Each volunteer collected the samples from the same stool sample using three different sampling kits (SK): a stool container (SK1); a flocked swab (SK2) and a cotton swab (SK3). The DNA was extracted using two isolation kits PowerLyzer PowerSoil DNA Isolation Kit (PS) and QIAamp DNA Stool Mini Kit (QS) (see Methods), totaling 96 samples for the analysis.

Evaluation of user acceptance of the sampling kits.

The participants were asked to select the best and the worst kit based on their ease of manipulation including the time spent using it. All 16 volunteers selected the stool container as the easiest to use and 13 out of 16 (81.25%) volunteers indicated the flocked swab as the worst sampling kit. We believe that the manipulation with cotton and flocked swabs is uncomfortable due to the small size and the necessity to insert the swab stick back into the tube without touching the tube wall. On the contrary, the stool container is easy to manipulate even for people with reduced motoric skills. In addition, the flocked swab is designed for sampling of liquid samples and the solid stool samples do not adhere on its synthetic fibers.

The effect of sampling and DNA isolation kits on the bacterial gDNA quality.

DNA yield, purity and integrity. Significantly higher DNA yields were obtained with the QS isolation kit, regardless of the sampling kit used ($q < 0.01$) (Fig. 1, Supplementary Table S1). The median values of the A260/A280 ratio (the measure of purity of DNA) were well within the expected range (1.8–2) and did not differ significantly between the DNA isolation kits or between the sampling kits (Fig. 1, Supplementary Table S1).

The DNA integrity was determined using the GQN measure (on a scale from 1 to 10; low GQN indicates strongly degraded gDNA sample) and the proportion of short fragments ($\leq$1500 bp; the larger the proportion the more degraded gDNA). We observed interaction effects of isolation and sampling kit for both DNA integrity measures. We found significantly lower proportion of short fragments when using the PS isolation kit (Fig. 1, Supplementary Table S1) and this difference was much larger when the stool container was used for sampling. There was no difference in GQN measure between the isolation kits when cotton or flocked swabs were used. However, for stool container samples, the QS kit provided much lower GQN values compared to the PS kit. These results point to worse DNA integrity for the QS isolation kit compared to the PS isolation kit mostly when stool container is used for sampling.

Presence of PCR inhibitors. The presence of PCR inhibitors in the samples decreases the sensitivity of the PCR reaction and even can lead to the impossibility of amplification of the selected region of 16S rRNA. It is usually measured by median efficiency values estimated from inhibition plots. Ideally, the efficiency should be 100%, meaning the template doubles in each cycle. Usually, the efficiency within 90–110% range is considered acceptable, where lower efficiency is caused by non-optimal reagent concentration or lower enzyme quality, while higher efficiency values are caused by the presence of PCR inhibitors. In our data, the efficiency values ranged from 96.7% to 114.0% (Fig. 1, Supplementary Table S2). In each of the isolation/sampling kit combinations, there were minimum two samples which exceeded the efficiency of 110%. The efficiency values of all isolation/sampling kit combinations, except for stool container samples after DNA isolation with the QS kit, were significantly increased compared to control samples without PCR inhibitors ($\text{efficiency}_{\text{med}} = 94.7\%$). No difference in efficiency values was observed between the isolation kits. The samples from stool containers (regardless of the isolation kit used) contained less PCR inhibitors in comparison to all other sampling/DNA isolation kit combinations (significantly lower efficiency, Supplementary Table S2). We hypothesize that this sampling kit effect is due to the sample dilution step prior to the DNA isolation step.

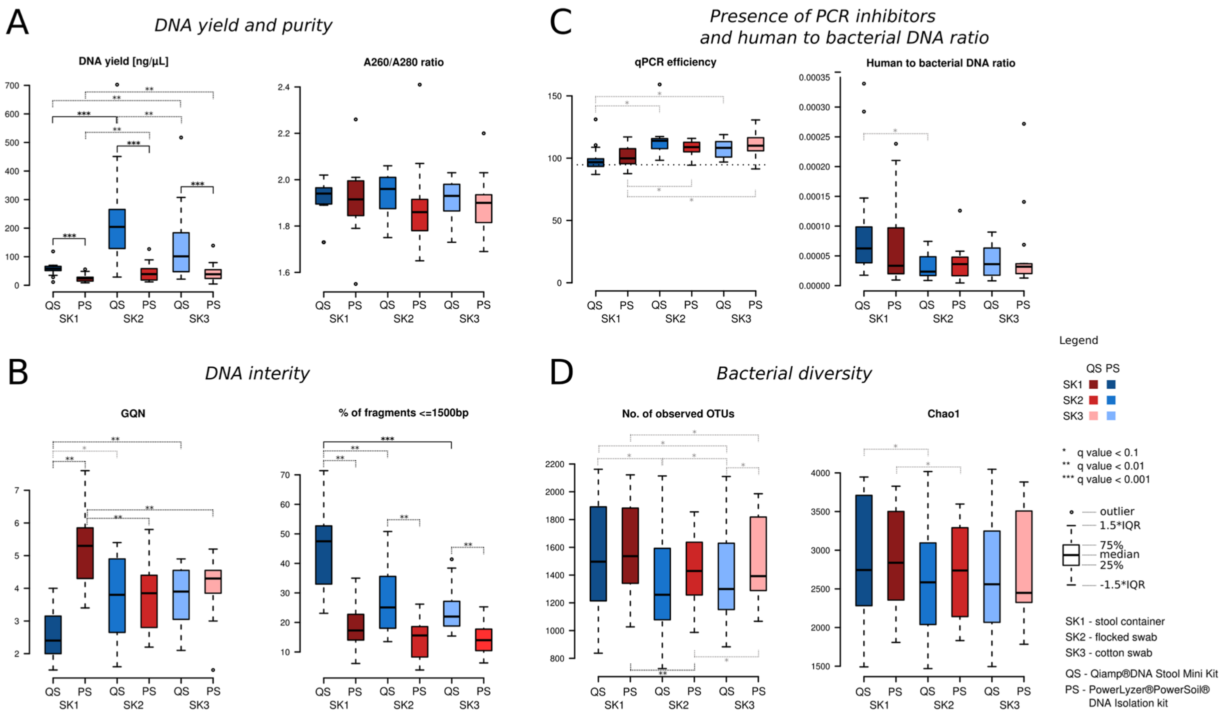


Figure 1. Comparison of sample DNA quality and diversity using different sampling and isolation kits. (**A**) DNA yield and purity comparison. ^{d}the samples were five times diluted prior the DNA extraction (see Methods); (**B**) DNA integrity comparison; (**C**) Presence of PCR inhibitors and human to bacterial DNA ratio comparison. Horizontal dotted line represents median efficiency value of the positive control; (**D**) Bacterial diversity comparison.

Human to bacterial DNA ratio. In all samples, the quantity of human DNA was lower than that of the bacterial DNA (ranging from 2947x to 221239x, median 29369x, see Fig. 1, Supplementary Table S2). No difference was found between sampling/isolation kit combinations in terms of human to bacterial DNA ratio, except for the increased ratio in the stool container compared to flocked swab samples after isolation with the QS kit (q = 0.03).

The effect of sampling and DNA isolation kits on bacterial diversity and composition.

Bacterial diversity. In total, 96 stool samples were sequenced. The number of reads after quality filtering and removal of chimeras ranged from 27680 to 67809, with median of 46192. We assessed the bacterial diversity using the number of observed OTUs and the Chao 1 diversity metric (Fig. 1, Supplementary Table S1). Overall, both diversity measures were independent of the DNA yield in all sampling/DNA isolation kit combinations.

While there was no difference in Chao 1 measure between the isolation kits, the number of observed OTUs was significantly increased after isolation with the PS kit, but only for cotton swab samples (q-value = 0.029). When comparing diversity measures between the sampling kits within each isolation kit separately, the stool container resulted in significantly higher number of observed OTUs in both DNA isolation kits (Fig. 1, Supplementary Table S1). In addition, we observed significantly higher number of OTUs in flocked swab samples compared to cotton swab samples after DNA isolation with the PS kit (q-value = 0.04) and significantly lower number of OTUs in flocked swab samples compared to cotton swab samples after DNA isolation with the QS kit (q-value = 0.09). For the Chao 1 diversity metric, significant differences were found in stool container samples compared to flocked swab samples in both PS and QS isolation kits (q = 0.04 and q = 0.09, respectively).

Bacterial composition. We identified 12,948 OTUs belonging to 13 bacterial phyla.

In order to quantify the effect of the sampling and isolation kits on bacterial composition, we performed mixed linear regression on each taxon that passed the filtering criteria (maximum abundance across all samples ≥1%) at all the seven taxonomical levels (phylum, class, order, family, genus, species, OTUs) separately. Interestingly, the proportion of taxa significantly affected by isolation or sampling kit differed between taxonomical levels (Fig. 2). The choice of sampling or DNA isolation kit affected 100% of taxa at phylum, class and order levels, and had decreasing trend from family to OTU level. The effects of sampling and isolation kits on the ten most abundant taxa at different taxa levels are summarized in Table 1 (see Supplementary Tables S3–S8 for complete results), the composition of significantly affected families is shown in Fig. 3. Overall, the choice of the isolation kit affected the abundance of more taxa than the choice of the sampling kit. In most of the cases where the taxa was affected by both factors, the p-values associated with the effect of the isolation kit were smaller than those of the sampling kit, indicating a more significant contribution of isolation kit to the overall model.

We hypothesized that the observed effect of the isolation kit was a result of different efficiency of the kit-specific bacterial cell walls lysis procedure. In this case, one of the kits would be more successful in isolating

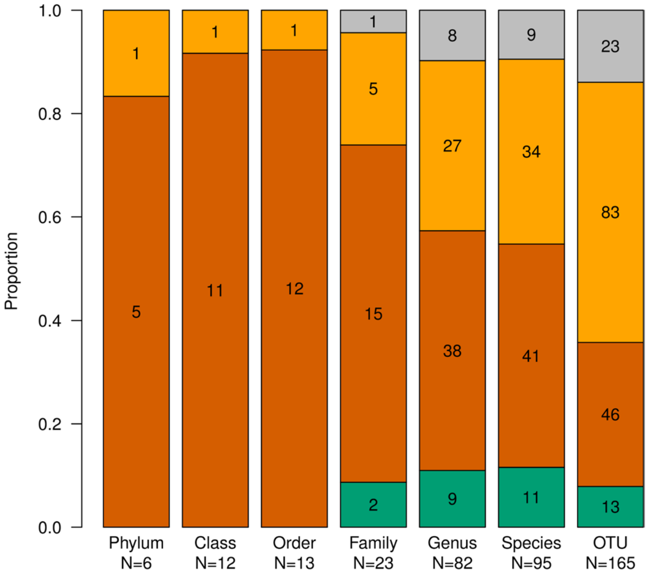


Figure 2. The proportion of taxa significantly affected by sampling or isolation kit at different taxonomical levels. Proportion of the tested taxa significantly affected by the sampling kit only (green), by the isolation kit only (dark yellow) and by both sampling and isolation kit (brick red). Grey indicates taxa not affected by sampling or isolation kit. The significance level was chosen at FDR < 10%, only taxa that met the selection criteria (maximum abundance >1%) were tested.

Gram-positive (G+) bacterial species. The Table 2 shows the numbers of significantly affected G+ taxa in all taxonomic levels and statistical pairwise comparison of their proportion after both isolation methods and all sampling methods. We found significantly higher proportions of G+ bacteria after the isolation using the PS kit at all the taxon levels. (96.4% to 100%, Table 2), compared to the QS isolation kit (G+ proportion varying from 0 to 44%). Similar observations were made for the effect of the sampling kit (Table 2), but this trend was not significant on any of the taxa levels except for the comparison of cotton swab (SK2) and stool container (SK1) on the genus level. We hypothesize that these differences are attributed to the dilution of the samples during the preprocessing steps specific to the stool container (see Methods for more details), resulting in lower sample density thus increasing the efficiency of the bead beating procedure. No difference in proportion of Gram-positive bacteria was found between flocked and cotton swabs. Figure 4 shows estimated effect sizes pairwise between the sampling kits on the genus level. Figure 5 visualizes bacteria with significant changes in abundance between isolation or sampling kits, with nodes colored according to Gram-positivity, where we can observe association of Gram-positive bacteria with the PS isolation kit.

Discussion

The gut microbiome seems to be crucial factor influencing human health and to date, a number of different diseases were correlated with microbiome dysbiosis. Understanding the true role of microbiome and fully comprehending its variability will require many cohort studies and, most probably, comparison of their results in large-scale meta-analyses. As with any other scientific domain, the incoherent methodological approaches constitute an important obstacle for such comparisons[44]. In an attempt to elucidate some of the factors determining the success of such studies, we focused on the effects of sampling and DNA extraction methods on a number of relevant variables from DNA integrity to final bacterial composition at different taxa levels. For this purpose, we selected sampling and DNA isolation kits that are the most common and accessible and hence are probably the most relevant for majority of cohort studies.

Our group of sixteen healthy volunteers used three different sampling kits – stool container, flocked swabs and cotton swabs. Without exception, the stool container was indicated as the most acceptable by the volunteers. Moreover, stool in the container can be easily diluted, homogenized and aliquoted for different analyses. Unfortunately, the stool container is inconvenient for sampling diarrhea or baby stool. Importantly, as we discuss below, the pre-processing specific to stool container samples influences both DNA quality and bacterial composition and these effects seem to interact with the DNA isolation kit.

For measuring the effect of different DNA extraction procedures, we used PowerLyzer PowerSoil DNA Isolation Kit (PS) and QIAamp DNA Stool Mini Kit (QS).

While the PS kit cell-wall lysis procedure is based on combination of bead-beating step and enzymatic lysis, the standard protocol of the QS kit comprises only enzymatic lysis. Considering the fact that the beat-beating step leads to higher DNA yield and higher number of observed OTUs from difficult-to-lyse bacteria, we added the bead-beating step also into the QS protocol, as commonly recommended[8,30,34,35,39].

DNA isolation by the QS kit resulted in significantly higher DNA yields compared to the PS kit (regardless of the sampling kit). Similar results were observed in other studies[30,32]. In agreement with previous studies[30,35,37], we found no significant correlation between DNA yield and alpha diversity.

Taxonomic level (# of all and significantly affected taxa)	Taxa (show ten most abundant)	q-value		Sign of the estimated effect size of the isolation or sampling kit				Relative abundance: total sum %	Gram stain
		isolation kit effect	sampling kit effect	PS to QS	SK2 to SK1	SK3 to SK1	SK3 to SK2		
Phylum All taxa: 14 Max >1% taxa: 6 Significantly affected taxa: 6 Isolation only: 1 Sampling only: 0 Both: 5	*Firmicutes*	**1.27E-15**	**4.43E-11**	+	−	−	+	68.3	G+
	Bacteroidetes	**3.42E-02**	**1.81E-02**	−	+	+	+	18.5	G−
	Actinobacteria	**3.67E-17**	**5.13E-04**	+	−	−	−	7.1	G+
	Proteobacteria	**5.05E-08**	**3.47E-04**	−	+	+	+	1.1	G−
	Verrucomicrobia	**1.69E-03**	**2.39E-03**	−	−	−	+	0.5	G−
	Tenericutes	**2.08E-03**	4.05E-01	−	−	−	−	0.1	G−
Class All taxa: 30 Max >1% taxa: 12 Significantly affected taxa: 12 Isolation only: 1 Sampling only: 0 Both: 11	*F; Clostridia*	**2.66E-15**	**9.45E-12**	+	−	−	+	61.2	G+
	B; Bacteroidia	**1.02E-02**	**2.15E-02**	−	+	+	+	18.5	G−
	A; Actinobacteria	**5.80E-14**	**3.04E-04**	+	−	−	−	4.6	G+
	F; Negativicutes	**2.58E-15**	**3.52E-04**	−	−	−	−	2.9	G−/var
	A; Coriobacteria	**5.80E-14**	**4.09E-03**	+	−	−	−	2	G+
	F; Erysipelotrichia	**2.66E-15**	**8.57E-05**	+	−	−	+	1.8	G+
	F; Bacilli	**2.58E-15**	**3.61E-05**	+	−	−	+	1	G+
	V; Verrucomicrobiae	**1.94E-03**	**1.33E-06**	−	−	−	+	0.5	G−
	P; Betaproteobacteria	**2.14E-05**	**1.49E-09**	−	+	+	+	0.4	G−
	P; Gammaproteobacteria	**3.67E-07**	**7.22E-03**	−	−	−	+	0.2	G−
Order All taxa: 49 Max >1% taxa: 13 Significantly affected taxa: 13 Isolation only: 1 Sampling only: 0 Both: 12	*F; Clostridiales*	**9.46E-13**	**1.37E-11**	+	−	−	+	61.2	G+
	B; Bacteroidales	**4.58E-03**	**1.69E-02**	−	+	+	+	18.5	G−
	A; Bifidobacteriales	**5.63E-12**	**2.30E-03**	+	−	−	−	4.5	G+
	F; Selenomonadales	**9.08E-17**	**2.42E-03**	−	−	−	+	2.9	G−/var
	A; Coriobacteriales	**5.73E-12**	**2.30E-02**	+	−	−	−	2	G+
	F; Erysipelotrichales	**3.05E-13**	**2.16E-04**	+	−	−	+	1.8	G+
	F; Lactobacillales	**3.05E-13**	**1.98E-04**	+	−	−	+	1	G+
	V; Verrucomicrobiales	**3.29E-04**	**1.22E-05**	−	−	−	+	0.5	G−
	P; Burkholderiales	**1.33E-05**	**1.70E-09**	−	+	+	+	0.4	G−
	T; Mollicutes	**1.04E-05**	**6.43E-05**	−	−	−	+	0.1	G−
Family All taxa: 85 Max >1% taxa: 23 Significantly affected taxa: 22 Isolation only: 5 Sampling only: 2 Both: 15	*F; Ruminococcaceae*	**9.20E-01**	**6.81E-13**	−	+	+	+	27.1	G+
	F; Lachnospiraceae	**5.68E-20**	**1.60E-03**	+	+	+	+	25.3	G+
	B; Bacteroidaceae	**7.55E-03**	**1.35E-02**	−	+	+	+	10.2	G−
	A; Bifidobacteriaceae	**5.90E-11**	**6.87E-03**	+	+	+	+	4.5	G+
	F; Veillonellaceae	**1.90E-12**	**1.55E-04**	−	+	+	+	2.4	G+
	A; Coriobacteriaceae	**7.68E-11**	**4.62E-02**	−	+	+	+	2	G+
	F; Erysipelotrichaceae	**1.74E-12**	**8.08E-04**	−	+	+	+	1.8	G+
	F; Christensenellaceae	**8.43E-01**	**1.24E-08**	−	+	+	+	1.4	G−
	B; Rikenellaceae	**5.90E-11**	7.24E-01	−	+	+	+	1.3	G−
	B; Porphyromonadaceae	**5.57E-04**	**4.03E-03**	−	+	+	+	1.1	G−
Genus All taxa: 277 Max >1% taxa: 82 Significantly affected taxa: 74 Isolation only: 27 Sampling only: 9 Both: 38	*B; Bacteroides*	**6.18E-03**	**1.54E-02**	−	+	+	+	10.2	G−
	F; Faecalibacterium	**1.37E-02**	**9.70E-05**	+	−	−	+	7.2	G+
	F; Blautia	**1.24E-24**	1.05E-01	+	−	−	+	5	G+
	A; Bifidobacterium	**1.48E-10**	**2.48E-02**	+	−	−	−	4.5	G+
	F; Subdoligranulum	**4.64E-03**	3.18E-01	−	−	−	+	3.7	G−
	F; Pseudobutyrivibrio	**9.63E-10**	6.32E-01	+	−	−	+	2.8	G−
	F; Dialister	**3.17E-09**	**5.86E-03**	−	−	−	+	2.2	G−
	F; Roseburia	**1.85E-02**	4.95E-01	+	−	−	+	1.5	G+
	A; Collinsella	**8.59E-05**	3.91E-01	+	+	−	−	1.4	G+
	F,Christensenellaceae R-7 group	6.06E-01	**1.50E-07**	−	−	−	−	1.4	G−

Table 1. Summary of taxa at all levels and detailed results for top 10 taxa significantly affected by sampling or DNA isolation kit. The significant q - values are shown in bold. SK1- stool container; SK2 – flocked swabs; SK3 - cotton swabs; PS – PowerLyzer PowerSoil DNA Isolation Kit; QS - QIAamp DNA Stool Mini Kit. All taxa – number of taxa found at the respective taxa level; Max >1% taxa – number of taxa that fulfilled the selection criteria for the analysis; Significantly affected taxa – the overall number of taxa at the respective taxa level affected by the isolation or sampling kit; Isolation only – number of taxa at the respective taxa level affected by the isolation kit only; Sampling only – number of taxa at the respective taxa level affected by the sampling kit only; Both – number of taxa at the respective taxa level affected by both sampling and isolation kit.

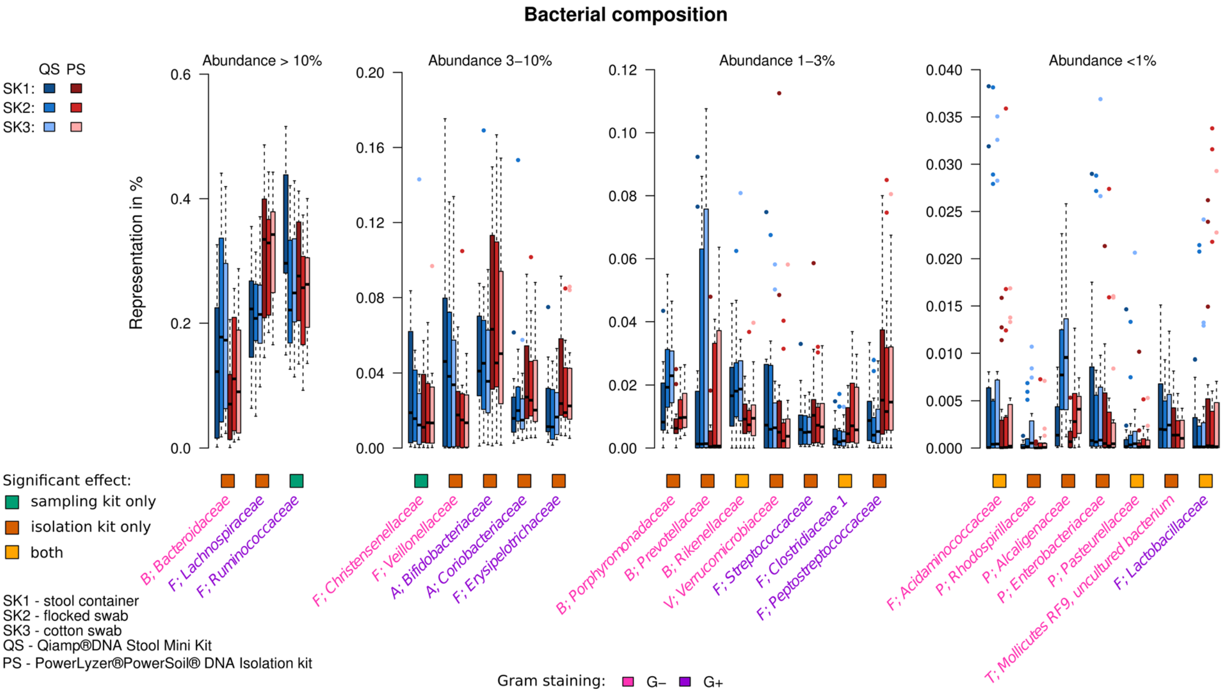


Figure 3. Distributions of relative abundances of significantly affected taxa at family level. Four graphs represent families divided according to third quartile of their abundance. Only taxa that passed the filtering criteria (maximum abundance >1%), significantly affected by isolation or sampling kit are shown. The colored squares below the graph indicate whether the family was affected significantly by the sampling kit only, the isolation kit only or both.

		Phylum		Class		Order		Family		Genus	
Sample groups	**Signif. more abundant**	**% of G+ phyla**	**q-val**	**% of G+ classes**	**q-val**	**% of G+ orders**	**q-val**	**% of G+ families**	**q-val**	**% of G+ geni**	**q-val**
PS to QS	in PS	100% (2/2)	**6.67E-02**	100% (5/5)	**2.71E-03**	100% (5/5)	**2.10E-03**	100% (8/8)	**1.98E-05**	96.4% (27/28)	**1.98E-05**
	in QS	0% (0/4)		0% (0/6)		0% (0/7)		0% (0/12)		44.1% (15/34)	
SK2 to SK1	in S1	66.7% (2/3)	4.00E-01	71.4% (5/7)	2.27E-01	62.5% (5/8)	2.27E-01	58.3% (7/12)	1.10E-01	80.8% (21/26)	**4.29E-02**
	in S2	0% (0/2)		0% (0/3)		0% (0/3)		0% (0/5)		44.4% (4/12)	
SK3 to SK1	in S1	66.7% (2/3)	4.00E-01	71.4% (5/7)	2.27E-01	62.5% (5/8)	2.27E-01	58.3% (7/12)	1.10E-01	80.0% (20/25)	1.10E-01
	in S3	0% (0/2)		0% (0/3)		0% (0/3)		0% (0/5)		38.5% (5/13)	
SK3 to SK2	in S3	25% (1/4)	5.37E-01	37.5% (3/8)	5.37E-01	33.3% (3/9)	5.37E-01	35.7% (5/14)	5.37E-01	60.7% (17/28)	5.37E-01
	in S2	50% (1/1)		100% (2/2)		100% (2/2)		66.7% (2/3)		80.0% (8/10)	

Table 2. Results of statistical testing of the proportion of G+ bacteria between significantly more abundant taxa within the selected isolation or sampling kit (pairwise). The significant q – values are shown in bold. SK1- stool container; SK2 – flocked swabs; SK3 – cotton swabs; PS – PowerLyzer PowerSoil DNA Isolation Kit; QS – QIAamp DNA Stool Mini Kit. Sample groups – which pairwise comparison was performed; Signif. more abundant – in which group the taxa were significantly more abundant; % of G+ taxa – proportion of G+ in the significantly more abundant taxa within the respective group and level.

On the other hand, the PS kit produced DNA of better integrity, even though in the PS protocol we applied more rigorous mechanical lysis (or higher speed of bead beating), which, according to the literature, should result in more degraded DNA[48]. We hypothesize that the observed differences might be caused by another factor, such as the type of the beads (0.1 mm glass in PS vs 0.1 mm zirconia in QS), the buffer composition, or the incubation temperature. Overall, for preparation of the shotgun libraries or sequencing using third generation of sequencers, we consider DNA integrity to be more important factor than the DNA yield, which favors PS kit over the QS kit.

To properly homogenize the samples from the stool container, we included a preprocessing procedure comprising five times dilution. This naturally resulted in lower yields of isolated DNA, but after adjustment for this dilution we obtained higher final DNA concentrations compared to undiluted stool samples from flocked and cotton swabs. It seems that the dilution step also affected the DNA integrity. Compared to the undiluted samples from flocked and cotton swabs, stool container samples resulted in less degraded DNA after isolation using the PS

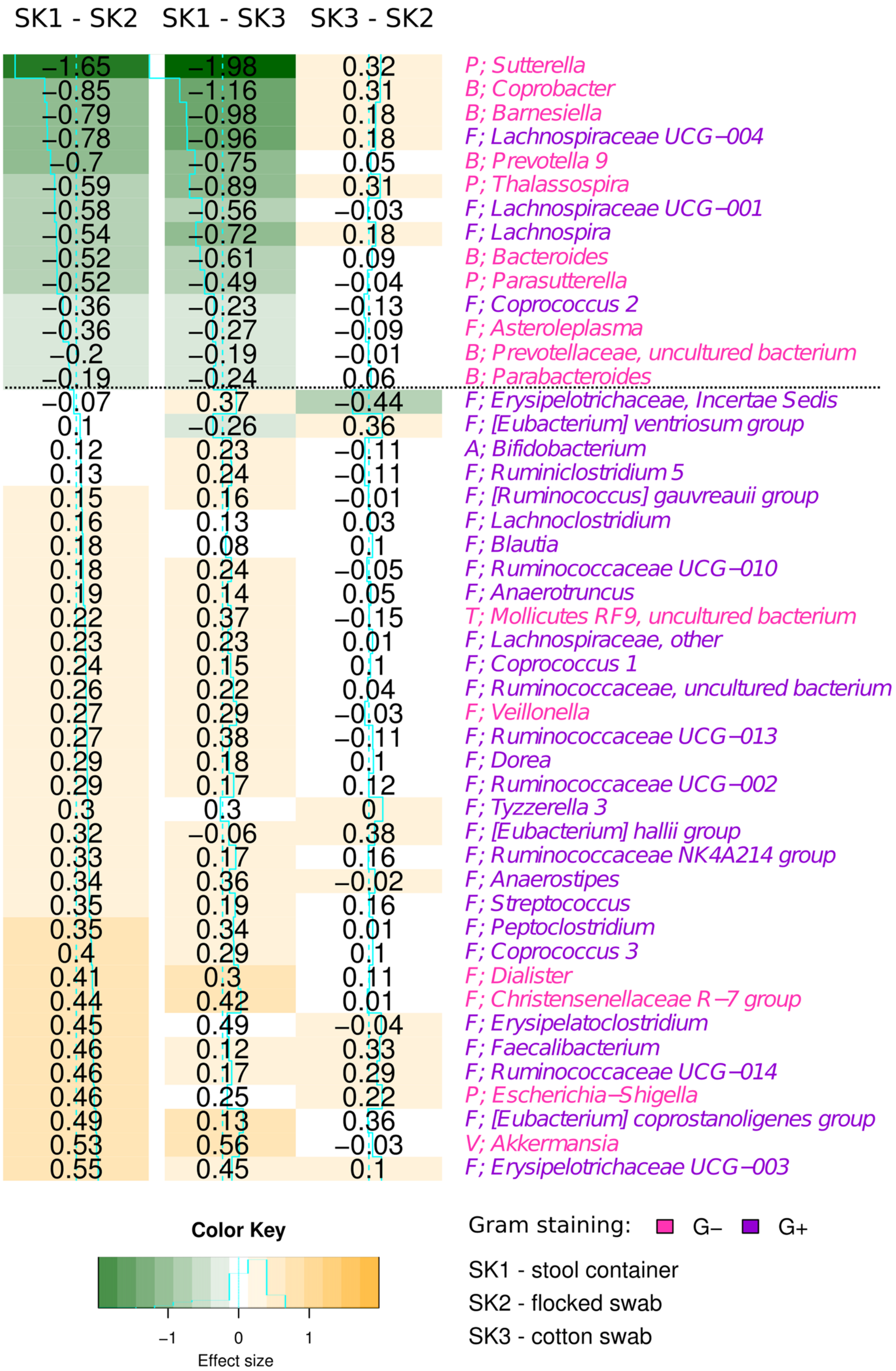


Figure 4. Comparison of sampling kits effects at genus level. Each column corresponds to a pair of sampling kits and each row corresponds to a specific bacteria genus. The values represent log fold changes of bacterial abundances (effect size) between the sampling kits, color coded from green (less abundant) to orange (more abundant). Only significantly affected taxa are shown.

kit and, in contrast, in more degraded DNA after isolation using the QS kit. Interestingly, two other independent studies, where different isolation kits were used, showed either a negative[34] or a positive[48] effect of sample dilution on the DNA integrity. This, together with our results leads us to conclude, that the effect of dilution step on DNA integrity is dependent on the isolation kit.

PCR inhibitors persisted in the DNA of the samples after isolation with both kits. Presence of PCR inhibitors could complicate the use of conventional molecular methods for the detection of low abundance or rare

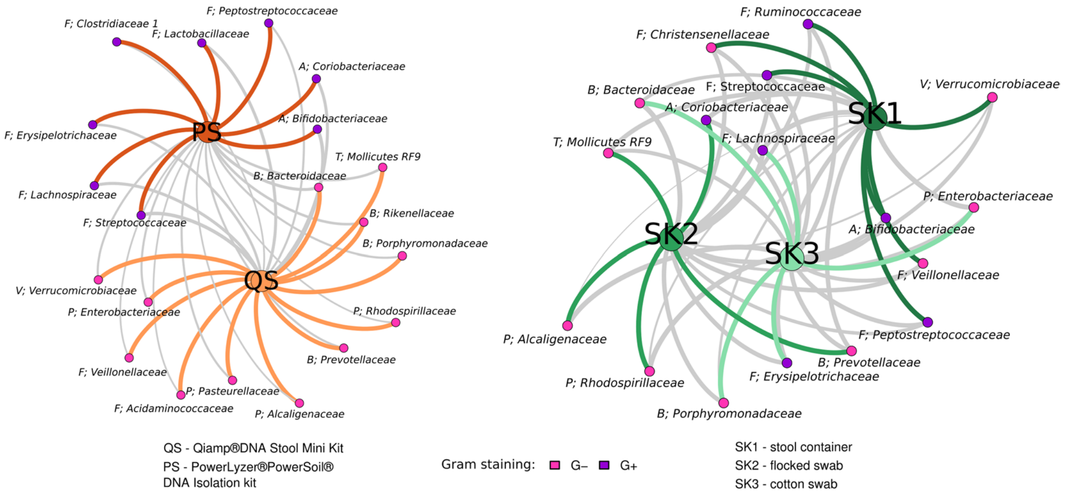


Figure 5. Association of bacterial families significantly differentially abundant between different sampling and isolation kits. The strength of the edges is weighted by relative abundance of taxa between the different kits (the stronger the edge, the larger the difference). Color-coding of the edges highlights taxa belonging to the same community, as detected by network modularity (see Methods for details). Grey edges represent connections between different communities.

pathogenic microorganisms[49,50]. The dilution of stool container samples prior to processing has led to significantly lower proportion of PCR inhibitors, hence for some applications, this approach might be preferred.

Both DNA extraction kits isolated preferentially bacterial DNA, independently on the sampling kit used and the amount of human DNA was negligible. From practical point of view, there is no superiority of any of the DNA isolation vs sampling kit combinations with respect to amount of residual human DNA. Some of the studies, however, use these kits to estimate the concentration of human DNA in stool samples as an indicator of inflammation that might predict onset of certain bowel diseases[51–55]. From this perspective, based on our results, we do not consider these kits eligible for human DNA quantification.

As for the alpha diversity, we observed increased number of OTUs after DNA isolation with the PS kit in all sampling kits, but the difference was significant only for cotton swab samples. We observed significant differences in number of OTUs between all sampling kits combinations, with the stool container resulting in the highest number of OTUs. We attribute the observed differences to higher effectivity of bead beating process in the less dense samples (the dilution preprocessing step used for the stool container). This is in contrast with the results of Santiago *et al.*[34], who report no changes in alpha diversity after sample dilution. In that study, however, a different isolation kit was used, so the results are not directly comparable.

The final bacterial composition was more affected by the choice of the DNA isolation kit than by the choice of the sampling kit. The preference of the PS isolation kit for Gram-positive bacteria was confirmed by statistical testing on all taxa levels and we believe that it is a result of more effective lysis of the Gram-positive cell wall bacteria when using the PS kit, despite the additional bead-beating step we introduced into the QS protocol. This is in agreement with previously published results[8,26]. It has to be taken into account, that Gram staining not always corresponds with the cell wall structure (e.g. *Pseudobutyrivibrio*[56] or *Deinococcus*[57], which is for many bacteria unknown. The efficiency of the lysis procedure can be as well influenced by atypical composition of the cell wall, presence of S-layer or capsules. The bacterial cell wall type also plays a role in the sampling effect: in our study it was associated with the dilution preprocessing step of the stool container, although less significantly.

There is a common belief that the effect of the individual is the most influential on the final bacterial composition[8,32]. Indeed, many metagenomic studies are reporting differences between groups of interest at the OTU level, where the effect of isolation and sampling is less important, as we showed in this study. However, some hypotheses are connecting particular disease with higher or lower bacterial abundance at the phylum or family level. An example is the commonly used *Bacteroidetes*/*Firmicutes* ratio[58–64]. Our results show, that this ratio is very dependent on both the selected DNA isolation method and sampling kit (dilution step). In our study, the PS kit and the dilution step (stool container) led to significantly higher proportion of e.g. *Firmicutes* (G+) and *Actinobacteria* (G+) and significantly lower proportion of *Proteobacteria* (G−) and *Bacteroidetes* (G−).

Another example of the cell wall structure effect is the Gram-positive genus *Blautia*. *Blautia* is a common and highly prevalent bacteria in the gastrointestinal tract, which is connected with healthy gut, since it is an effective short-chain fatty acid producer[65,66]. Lower abundance of *Blautia* in the gut is associated with many diseases[66–73]. In our study, *Blautia* was bacteria the most significantly affected by DNA isolation (across all the taxonomic levels). Similar observations were also described as the effect of isolation in other studies[26,34].

The sampling kit (dilution effect) influenced most significantly the abundance of genus *Sutterella*, bacteria correlated with many diseases such as celiac diseases[67], Down syndrome[74], autism[75] or irritable bowel syndrome[76]. Clearly, the dilution step represents an important batch effect, which raises a question, whether it is related only to the artificial dilution, or this effect could also be observed in diarrheic samples. The effect of stool consistency was described previously as an important factor[12,77,78] influencing the bacterial composition, but this effect was not connected with effect of higher water content (dilution), rather with the transit time. As previously recommended[77], we also suggest to control for the stool consistency as a potential confounding factor to avoid the effect of sample water content in this kind of studies, especially if one of the illness symptoms is diarrhea.

Despite the fact that the significance of the sampling and isolation dependent batch effects is repeatedly reported, no systematic study of these effects was performed yet on samples from larger numbers of individuals. Efforts for standardization of laboratory practices in metagenomics have been made in large international projects such as Metagenomic Research Group (MGRG), Genomic Standard Consortium (GSC), The Microbiome Quality Control Project (MBQC) and International Human Microbiome Standards (IHMS). IHMS recommends a procedure for fecal sample DNA extraction, based on study of Costea *et al*., where 21 extraction protocols were compared, including protocols similar to ours – protocol 3 (with PowerLyzer PowerSoil DNA Isolation Kit) and 11 (with QIAamp DNA Stool Mini Kit and bead beating step)[39]. They selected the protocol with QIAamp DNA Stool Mini Kit as the best choice for its accuracy and reproducibility. In contrast to our results, both protocol 3 and 11, provide good lysis of Gram-positive bacteria, but protocol 3 was excluded for insufficient DNA quality. The main difference between the studies is that the Costea study was based on the results of whole metagenomics sequencing and only compared bacterial composition annotated at the species level.

All these above mentioned studies and our results confirm that meta-analytical studies are extremely challenging due to the many sources of batch effects that need to be accounted for. Incorporation of a standardized mock community to the sequencing workflow, followed by normalization of the results to these reference values could be solution in future. The increased cost per run and slightly more complex library preparation is a small price to pay for robustness, consistency and comparability of results.

Conclusions

We performed systematic study of effects of DNA isolation and sampling kit on DNA quality and bacterial composition based on sequencing of gene for 16S rRNA on a the largest number of individuals to day (96 samples from 16 individuals).

We found significant effect of both DNA isolation and sampling kits on DNA purity, DNA integrity, alpha diversity and bacterial composition. Overall, the DNA isolation effect was stronger than that of the sampling kit. Interestingly the proportion of taxa affected by isolation or sampling was decreasing with decreasing taxonomical level.

We confirmed previously reported effect of DNA isolation kit on bacterial composition due to bacterial cell wall structure, namely the better efficacy of The PowerLyzer PowerSoil DNA Isolation Kit in lysis of Gram-positive bacteria. In addition, we report that the dilution pre-processing step of the stool container samples favored Gram-positive bacteria, although mostly at the genus level.

Both the choice of isolation and sampling kits significantly affected the *Firmicutes* to *Bacteroidetes* ratio. We conclude that the choice of DNA isolation and sampling kit (dilution step, and by extension the stool consistency) is an important batch effect that has to be taken into account mainly when comparing results between studies.

Methods

Sample collection.

Stool samples were collected from a group of 16 volunteers. The subjects were 23–65 years old with an average age of 40.9 and none of them suffered from diarrhea during sample collection. Stool samples were collected at home. Volunteers received three stool sampling kits: sampling kit 1 (SK1) comprising 1x stool container (FL Medical, Italy); sampling kit 2 (SK2) comprising 2x flocked swabs (Copan, Italy) and sampling kit 3 (SK3) comprising 2x cotton swabs (SceneSafe, Great Britain). Sampling kits also contained disposable gloves and hand and surface disinfectant wipes for more convenient sampling. Each volunteer was instructed to collect all the samples from the same stool and from the same spot. Stool samples were then stored in a freezer at −20 °C overnight to freeze completely and the next day were transported on ice buckets to the laboratory, where they were stored at −20 °C prior to processing. Each group of samples was processed at the same time and by the same person. Participants filled out a brief questionnaire about satisfaction with individual sampling kits after stool sample collection. The study design is summarized in Fig. 6.

This study was carried out in accordance with the recommendations of the ELSPAC Steering Committee of Masaryk University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocols were approved by the ELSPAC Steering Committee of Masaryk University.

DNA extraction.

Stool in the stool container (SK1) was diluted 5x with molecular grade water and homogenized by vortexing with Zirconia beads 2.3 mm (BioSpec, USA) to receive identical aliquots. This step is not necessary for the swabs, since each swab serves as an aliquot itself. Stool suspension (250 μl) was used for DNA extractions. Flocked swabs (SK2) and cotton swabs (SK3) were transferred into 2 ml tubes to be prepared for subsequent DNA extraction. DNA extractions were performed using a PowerLyzer PowerSoil DNA Isolation Kit (Mo Bio, USA) (PS) and QIAamp DNA Stool Mini Kit (Qiagen, USA) (QS) according to the manufacturer's instructions.
Deviations from PS protocol:

- 750 μl of Bead Solution and 60 μl of C1 Solution were added to swab samples (SK2 and SK3) after defrosting. Samples were thoroughly vortexed and centrifuged for 4 min at 36,220 RCF. The swabs were then removed. Next, the samples were homogenized using the FastPrep-24 (MP Biomedicals, USA) 45 s 6.5 m/s.

Deviations from QS protocol:

- A homogenization step with 0.1 mm zirconia beads (BioSpec, USA) was added to the protocol after the third step (i.e. after the suspension was heated for 5 min at 95 °C).

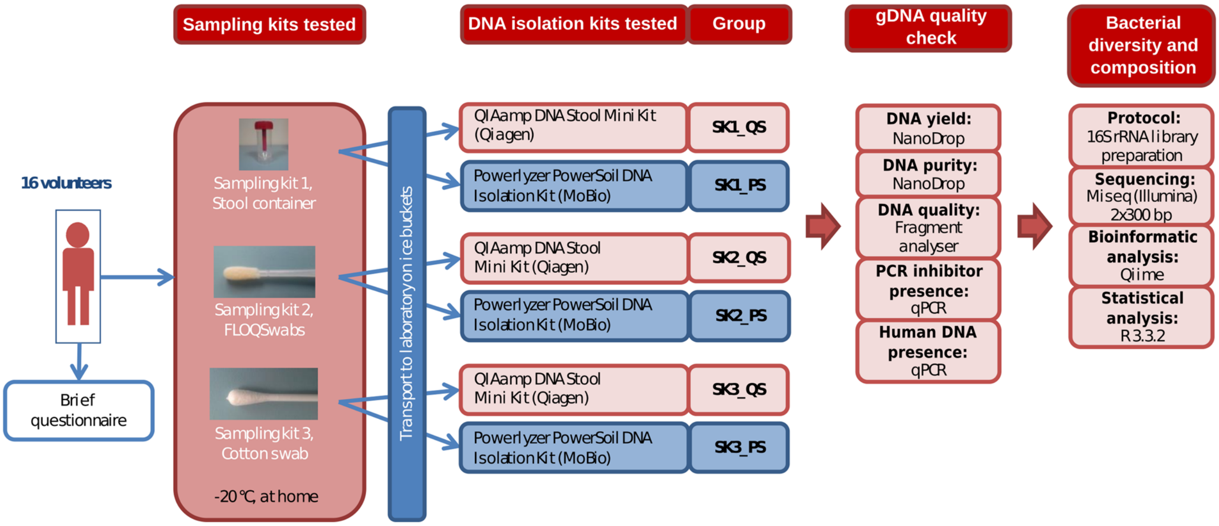

Sampling kits tested
DNA isolation kits tested
Group
gDNA quality check
Bacterial diversity and composition
16 volunteers
Brief questionnaire
Sampling kit 1, Stool container
Sampling kit 2, FLOQSwabs
Sampling kit 3, Cotton swab
-20 °C, at home
Transport to laboratory on ice buckets
QIAamp DNA Stool Mini Kit (Qiagen)
SK1_QS
PowerIyzer PowerSoil DNA Isolation Kit (MoBio)
SK1_PS
QIAamp DNA Stool Mini Kit (Qiagen)
SK2_QS
PowerIyzer PowerSoil DNA Isolation Kit (MoBio)
SK2_PS
QIAamp DNA Stool Mini Kit (Qiagen)
SK3_QS
PowerIyzer PowerSoil DNA Isolation Kit (MoBio)
SK3_PS
DNA yield: NanoDrop
DNA purity: NanoDrop
DNA quality: Fragment analyser
PCR inhibitor presence: qPCR
Human DNA presence: qPCR
Protocol: 16S rRNA library preparation
Sequencing: Miseq (Illumina) 2x300 bp
Bioinformatic analysis: Qiime
Statistical analysis: R 3.3.2

Figure 6. Study design. Flowchart summarizing the study design and methods used.

Target region/gene	Amplicon size	Primer name	Primer Sequences (5′ → 3′)	Cycling conditions						Reference
16S rRNA gene (bacterial DNA)	146 bp	q16S-univF	GTGSTGCAYGGYTGTCGTCA	95 °C	45x	95 °C	53 °C	72 °C		Maeda *et al.*[90]
		q16S-univR	ACGTCRTCCMCACCTTCCTC	10 min		20 s	30 s	20 s	—	
GAPDH (human DNA)	74 bp		TGCACCACCAACTGCTTAGC	95 °C	40x	95 °C	65 °C			Vandesompele *et al.*[91]
			GGCATGGACTGTGGTCATGAG	10 min		10 s	60 s	—	—	
V3/V4 16S rRNA gene (library preparation)	~460 bp	s16S_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-InnerTag-CCTACGGGNGGCWGCAG	95 °C	25x	95 °C	55 °C	72 °C	72 °C	Klindworth *et al.*[79]
		s16S_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-InnerTag-GACTACHVGGGTATCTAATCC	3 min		30 s	30 s	30 s	5 min	

Table 3. Primers and cycling conditions used in this study.

- 1.4 µl Buffer ASL was added to swab samples (SK2 and SK3) after defrosting. Samples were vortexed continuously for 1 min and the suspension was heated for 5 min at 70 °C. Next, the samples were homogenized using the FastPrep-24 (MP Biomedicals, USA) 45 s 5.5 m/s.

Evaluation of DNA yield, purity and quality. The final yield of extracted DNA was determined spectrophotometrically using theNanoDropND-1000 (Thermo Fisher SCIENTIFIC, USA). The purity of extracted DNA was indicated by an A260/A280 nm ratio. The quality of extracted DNA was assessed using the Fragment Analyzer (Advanced Analytical Technologies, USA) and High Sensitivity Genomic DNA Analysis Kit (Advanced Analytical Technologies, USA). The percentage of short fragments (≤1,500 bp) and Genomic Quality Number (GQN threshold of 10,000 bp) were calculated by PROSize 2.0 (Advanced Analytical Technologies, USA). Extracted DNA from each sample was diluted approximately to 5 ng/µl, aliquoted and stored at −20 °C. Aliquots were subsequently used in all further methods as starting material.

Presence of PCR inhibitors after different DNA extractions. The presence of inhibitors was tested by qPCR. A primer pair specific for the conservative regions of 16S rRNA gene (Table 3) was used. qPCR was performed on the TOptical Thermocycler (Analytik Jena - Biometra, Ireland) using a KAPA SYBR FAST qPCR Kit (Kapa Biosystems, USA). Cycling conditions are described in Table 2. Melting temperature was determined after PCR to verify the correctness of each PCR product. Extracted DNA from four different isolates of *Escherichia coli* DH10B served as a positive control without PCR inhibitors. Each extracted DNA from sample and positive control (concentration approximately 5 ng/µl) was diluted three times (10x, 100x, 1,000x). The subsequent qPCR reactions were performed using both diluted and undiluted samples. Inhibition plots were created from Ct values and efficiency ($=10^{(-1/\text{slope})}-1$) was calculated for each sample and positive control.

Proportion of human DNA to bacterial DNA after different DNA extractions. The ratio of human and bacterial DNA in samples was tested by qPCR. Bacterial DNA was assessed using a primer pair specific for the conservative regions of 16S rRNA gene and human DNA using a primer pair specific for protein kinase (Table 3). qPCRwas performed on the TOptical Thermocycler (Analytik Jena - Biometra, Ireland) with KAPA SYBR FAST qPCR Kit (Kapa Biosystems, USA). Cycling conditions are described in Table 3. Melting temperature was determined after PCR to verify the correctness of each PCR product. The amount of human DNA to bacterial DNA was calculated as $2^{\Delta Ct}$. Ct value of 40 was used for all samples under the limit of detection.

PCR amplification and Illumina library preparation. Extracted DNA was used as a template in amplicon PCR to target the hypervariable V3 and V4 regions of the bacterial 16S rRNA gene. The 16S metagenomics library was prepared according to the Illumina 16S Metagenomic sequencing Library Preparation protocol with some deviations described below (for workflow diagram see Supplementary Fig. S1). Each PCR was performed in triplicate, with the primer pair consisting of Illumina overhang nucleotide sequences, an inner tag and gene-specific sequences[79]. The Illumina overhang served to ligate the Illumina index and adapter. Each inner tag, i.e. a unique sequence of 7–9 bp, was designed to differentiate samples into groups. Primer sequences and PCR cycling conditions are summarized in Table 3. After PCR amplification, triplicates were pooled and the amplified PCR products were determined by gel electrophoresis. PCR clean-up was performed with Agencourt AMPure XP beads (Beckman Coulter Genomics, USA). Samples with different inner tags were equimolarly pooled based on fluorometrically measured concentration using Qubit dsDNA HS Assay Kit (Invitrogen, USA) and microplate reader Synergy Mx (BioTek, USA). Pools were used as a template for a second PCR with Nextera XT indexes (Illumina, USA). Differently indexed samples were quantified using the KAPA Library Quantification Complete Kit (Kapa Biosystems, USA) and equimolarly pooled according to the measured concentration. The prepared library was checked with a 2100 Bioanalyzer Instrument (Agilent Technologies, USA) and concentration was measured with qPCR shortly before sequencing. The library was diluted to a final concentration of 8 pM and 20% of PhiX DNA (Illumina, USA) was added. Sequencing was performed with the Miseq reagent kit V3 using a MiSeq. 2000 instrument according to the manufacturer's instructions (Illumina, USA).

Data analysis. Forward and reverse pair-end reads, that fulfilled the condition of both quality and length filtering, were merged using the fastq-join method within the join_pair_ends.py command in QIIME 1.9.1[80]. Data were demultiplexed and barcodes and primers were trimmed using package Biostrings[81] in R 3.3.2[82]. Operational taxonomic units (OTUs) were constructed by binding sequences into clusters of greater than 97% sequence similarity using QIIME. In the next step, chimeras were detected on the set of representative sequences of each OTU with UCHIME in USEARCH v6.1.544[83]. These chimera OTUs were subsequently excluded from the analysis. Taxonomy was assigned to each OTU based on SILVA 123 reference database[84]. The observed species metric and the Chao1 index were used to estimate alpha diversity for each sample in QIIME. Beta diversity was computed in QIIME using both weighted and unweighted UniFrac metrics[85]. All statistical analysis was performed in R 3.3.2[82].

The data were treated as compositional (proportions of total read count in each sample, non-rarefied) and prior to all statistical analyses were transformed using centered log-ratio transformation[86]. The analyses were performed on each of the seven taxonomy levels (Phylum, Class, Order, Family, Genus, Species and OTUs) separately and the resulting p-values were adjusted for multiple hypothesis testing using Benjamini-Hochberg procedure. Results were considered significant at FDR = 10%. The adjusted p-values are referred to as q-values.

To estimate the effects of isolation and sampling kits on bacterial composition while accounting for repeated measurements (effect of individual), we applied linear mixed model with sampling and izolation kits as fixed effects and individual as random effect (intercept). Log-likelihood test was performed to detect significance of each of the fixed effects – each time we compared the full model to the model without the fixed effect of interest.

A non-parametric Wilcoxon paired test, was used for comparison of effect of isolation kits on DNA quality. We used Spearman's rank order correlation coefficient to discover the strength of the link between the number of observed species and DNA concentration.

Bipartite networks were used to visualize the influence of different kits on detection of Gram-positive and Gram-negative bacteria. These networks were reconstructed according to Sedlar *et al.*[87] using R 3.3.2 and visualized in Gephi 0.9.2[88,89]. Communities within networks were extracted using modularity optimization criterion[88].

Data Availability

Sequencing data were uploaded to the European Nucleotide Archive under accession number PRJEB24411. Read counts per sample at different taxa levels and sample information table are available in Supplementary Files S9–S11.

References

1. Suau, A. *et al*. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl. Environ. Microbiol.* **65**, 4799–807 (1999).
2. Zoetendal, E. G. *et al*. Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Appl. Environ. Microbiol.* **68**, 3401–7 (2002).
3. Russell, S. L. *et al*. Perinatal antibiotic treatment affects murine microbiota, immune responses and allergic asthma. *Gut Microbes* **4**, 158–64 (2013).
4. Jandhyala, S. M. *et al*. Role of the normal gut microbiota. *World J. Gastroenterol.* **21**, 8787–803 (2015).
5. Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G. & De La Cochetiere, M. F. Development of intestinal microbiota in infants and its impact on health. *Trends in Microbiology* **21**, 167–173 (2013).
6. Underwood, M. A. Intestinal dysbiosis: Novel mechanisms by which gut microbes trigger and prevent disease. *Prev. Med. (Baltim).* **65**, 133–137 (2014).
7. Zhang, Y.-J. *et al*. Impacts of gut bacteria on human health and diseases. *Int. J. Mol. Sci.* **16**, 7493–519 (2015).
8. Mackenzie, B. W., Waite, D. W. & Taylor, M. W. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Front. Microbiol.* **6**, 130 (2015).
9. Yatsunenko, T. *et al*. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
10. Abeles, S. R. *et al*. Microbial diversity in individuals and their household contacts following typical antibiotic courses. *Microbiome* **4**, 39 (2016).

11. Korpela, K. & de Vos, W. Antibiotic use in childhood alters the gut microbiota and predisposes to overweight. *Microb. Cell* **3**, 296–298 (2016).
12. Zhernakova, A. *et al*. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–9 (2016).
13. Graf, D. *et al*. Contribution of diet to the composition of the human gut microbiota. *Microb. Ecol. Health Dis.* **26**, 26164 (2015).
14. Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4**, 15 (2016).
15. Claus, S. P., Guillou, H. & Ellero-Simatos, S. The gut microbiota: a major player in the toxicity of environmental pollutants? *npj Biofilms Microbiomes* **2**, 16003 (2016).
16. Madan, J. C., Farzan, S. F., Hibberd, P. L. & Karagas, M. R. Normal neonatal microbiome variation in relation to environmental factors, infection and allergy. *Curr. Opin. Pediatr.* **24**, 753–9 (2012).
17. Schultze, A. *et al*. Comparison of stool collection on site versus at home in a population-based study. *Bundesgesundheitsblatt - Gesundheitsforsch. - Gesundheitsschutz* **57**, 1264–1269 (2014).
18. Feigelson, H. S. *et al*. Feasibility of self-collection of fecal specimens by randomly sampled women for health-related studies of the gut microbiome. *BMC Res. Notes* **7**, 204 (2014).
19. Loftfield, E. *et al*. Comparison of collection methods for fecal samples for discovery metabolomics in epidemiologic studies. *Cancer Epidemiol. Biomarkers Prev.* **25**, 1483–1490 (2016).
20. Tedjo, D. I. *et al*. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS One* **10**, e0126685 (2015).
21. Mathay, C. *et al*. Method Optimization for Fecal Sample Collection and Fecal DNA Extraction. *Biopreserv. Biobank.* **13**, 79–93 (2015).
22. Panek, M. *et al*. Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. *Sci. Rep.* **8**, 5143 (2018).
23. Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R. & Fierer, N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol. Lett.* **307**, 80–86 (2010).
24. Cardona. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol*. **12**, 158 (2012).
25. Gorzelak, M. A. *et al*. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS One* **10**, e0134802 (2015).
26. Maukonen, J., Simões, C. & Saarela, M. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiol. Ecol.* **79**, 697–708 (2012).
27. Hill, C. J. *et al*. Effect of room temperature transport vials on DNA quality and phylogenetic composition of faecal microbiota of elderly adults and infants. *Microbiome* **4**, 19 (2016).
28. Choo, J. M., Leong, L. E. X. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5**, 16350 (2015).
29. Kim, D. *et al*. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52 (2017).
30. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. *PLoS One* **7**, e33865 (2012).
31. Janabi, A. H. D., Kerkhof, L. J., McGuinness, L. R., Biddle, A. S. & McKeever, K. H. Comparison of a modified phenol/chloroform and commercial-kit methods for extracting DNA from horse fecal material. *J. Microbiol. Methods* **129**, 14–19 (2016).
32. Kennedy, N. A. *et al*. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9**, e88982 (2014).
33. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Krogfelt, K. A. Optimising bacterial DNA extraction from faecal samples: comparison of three methods. *Open Microbiol. J.* **5**, 14–7 (2011).
34. Santiago, A. *et al*. Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14**, 112 (2014).
35. Gerasimidis, K. *et al*. The effect of DNA extraction methodology on gut microbiota research applications. *BMC Res. Notes* **9**, 365 (2016).
36. Claassen, S. *et al*. A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples. *J. Microbiol. Methods* **94**, 103–110 (2013).
37. Salonen, A. *et al*. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81**, 127–134 (2010).
38. Lim, M. Y., Song, E.-J., Kim, S. H., Lee, J. & Nam, Y.-D. Comparison of DNA extraction methods for human gut microbial community profiling. *Syst. Appl. Microbiol.* **41**, 151–157 (2018).
39. Costea, P. I. *et al*. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
40. Walker, A. W. *et al*. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**, 26 (2015).
41. Cruaud, P. *et al*. Influence of DNA Extraction Method, 16S rRNA Targeted Hypervariable Regions, and Sample Origin on Microbial Diversity Detected by 454 Pyrosequencing in Marine Chemosynthetic Ecosystems. *Appl. Environ. Microbiol.* **80**, 4626–4639 (2014).
42. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS One* **6**, e27310 (2011).
43. Clooney, A. G. *et al*. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLoS One* **11**, e0148028 (2016).
44. Lozupone, C. A. *et al*. Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
45. Fu, B. C. *et al*. Characterization of the gut microbiome in epidemiologic studies: the multiethnic cohort experience. *Ann. Epidemiol.* **26**, 373–379 (2016).
46. Hsieh, Y.-H. *et al*. Impact of Different Fecal Processing Methods on Assessments of Bacterial Diversity in the Human Intestine. *Front. Microbiol.* **7**, 1643 (2016).
47. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–30 (2012).
48. Bürgmann, H., Pesaro, M., Widmer, F. & Zeyer, J. A strategy for optimizing quality and quantity of DNA extracted from soil. *J. Microbiol. Methods* **45**, 7–20 (2001).
49. Schrader, C., Schielke, A., Ellerbroek, L. & Johne, R. PCR inhibitors - occurrence, properties and removal. *J. Appl. Microbiol.* **113**, 1014–1026 (2012).
50. Oikarinen, S. *et al*. PCR inhibition in stool samples in relation to age of infants. *J. Clin. Virol.* **44**, 211–214 (2009).
51. Lewis, J. D. *et al*. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500 (2015).
52. TEIXEIRA, Y. *et al*. Human dna quantification in the stools of patients with colorectal cancer. *Arq. Gastroenterol.* **52**, 293–298 (2015).
53. Varela, E. *et al*. Faecal DNA and calprotectin as biomarkers of acute intestinal toxicity in patients undergoing pelvic radiotherapy. *Aliment. Pharmacol. Ther.* **30**, 175–85 (2009).
54. Zou, H., Harrington, J. J., Klatt, K. K. & Ahlquist, D. A. A sensitive method to quantify human long DNA in stool: relevance to colorectal cancer screening. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1115–9 (2006).

55. Klaassen, C. H. W. *et al*. Quantification of human DNA in feces as a diagnostic test for the presence of colorectal cancer. *Clin. Chem.* **49**, 1185–7 (2003).
56. Hespell, R. B., Kato, K. & Costerton, J. W. Characterization of the cell wall of Butyrivibrio species. *Can. J. Microbiol.* **39**, 912–921 (1993).
57. Thompson, B. G. & Murray, R. G. Isolation and characterization of the plasma membrane and the outer membrane of Deinococcus radiodurans strain Sark. *Can. J. Microbiol.* **27**, 729–34 (1981).
58. Karlsson, F. H. *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
59. Li, E. *et al*. Inflammatory Bowel Diseases Phenotype, C. difficile and NOD2 Genotype Are Associated with Shifts in Human Ileum Associated Microbial Composition. *PLoS One* **7**, e26284 (2012).
60. Gevers, D. *et al*. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
61. Gao, Z., Guo, B., Gao, R., Zhu, Q. & Qin, H. Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* **6**, 20 (2015).
62. Turnbaugh, P. J. *et al*. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–31 (2006).
63. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
64. Larsen, N. *et al*. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5**, e9085 (2010).
65. Tanaka, S., Yamamoto, K., Yamada, K., Furuya, K. & Uyeno, Y. Relationship of Enhanced Butyrate Production by Colonic Butyrate-Producing Bacteria to Immunomodulatory Effects in Normal Mice Fed an Insoluble Fraction of Brassica rapa L. *Appl. Environ. Microbiol.* **82**, 2693–9 (2016).
66. Murri, M. *et al*. Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Med.* **11**, 46 (2013).
67. Cheng, J. *et al*. Duodenal microbiota composition and mucosal homeostasis in pediatric celiac disease. *BMC Gastroenterol.* **13**, 113 (2013).
68. Chen, W., Liu, F., Ling, Z., Tong, X. & Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One* **7**, e39743 (2012).
69. Hong, P.-Y., Croix, J. A., Greenberg, E., Gaskins, H. R. & Mackie, R. I. Pyrosequencing-based analysis of the mucosal microbiota in healthy individuals reveals ubiquitous bacterial groups and micro-heterogeneity. *PLoS One* **6**, e25042 (2011).
70. Bajaj, J. S. *et al*. Colonic mucosal microbiome differs from stool microbiome in cirrhosis and hepatic encephalopathy and is linked to cognition and inflammation. *Am. J. Physiol. Gastrointest. Liver Physiol.* **303**, G675–85 (2012).
71. Schnabl, B. & Brenner, D. A. Interactions between the intestinal microbiome and liver diseases. *Gastroenterology* **146**, 1513–1524 (2014).
72. Org, E. *et al*. Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the METSIM cohort. *Genome Biol.* **18**, 70 (2017).
73. Lippert, K. *et al*. Gut microbiota dysbiosis associated with glucose metabolism disorders and the metabolic syndrome in older adults. *Benef. Microbes* 1–12, https://doi.org/10.3920/BM2016.0184 (2017).
74. Biagi, E. *et al*. Gut microbiome in Down syndrome. *PLoS One* **9**, e112023 (2014).
75. Wang, L. *et al*. Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder. *Mol. Autism* **4**, 42 (2013).
76. Mukhopadhya, I. *et al*. A Comprehensive Evaluation of Colonic Mucosal Isolates of Sutterella wadsworthensis from Inflammatory Bowel Disease. *PLoS One* **6**, e27076 (2011).
77. Vandeputte, D. *et al*. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2016).
78. Falony, G. *et al*. Population-level analysis of gut microbiome variation. *Science* **352**, 560–4 (2016).
79. Klindworth, A. *et al*. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
80. Chen, H. M. & Lifschitz, C. H. Preparation of fecal samples for assay of volatile fatty acids by gas-liquid chromatography and high-performance liquid chromatography. *Clin. Chem.* **35**, 74–76 (1989).
81. Pagés, H., Aboyout, P., Gentleman, R. & Biostrings, D. S. String objects representing biological sequences, and matching algorithms (2016).
82. R Core Team (2016). R: *A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2016).
83. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
84. Pruesse, E. *et al*. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
85. Lozupone, C. & Knight, R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
86. Aitchison, J. (John). *The statistical analysis of compositional data*. (Chapman and Hall, 1986).
87. Sedlar, K., Videnska, P., Skutkova, H., Rychlik, I. & Provaznik, I. Bipartite graphs for visualization analysis of microbiome data. *Evol. Bioinforma.* **12** (2016).
88. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
89. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs.
90. Maeda, H. *et al*. Quantitative real-time PCR using TaqMan and SYBR Green for Actinobacillus actinomycetemcomitans, Porphyromonas gingivalis, Prevotella intermedia, tetQ gene and total bacteria. *FEMS Immunol. Med. Microbiol.* **39** (2003).
91. Vandesompele, J. *et al*. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, RESEARCH0034 (2002).

Acknowledgements

We acknowledge the help of the volunteers enrolled in this study. This study was funded by the Ministry of Education, Youth and Sports of the Czech Republic and European Structural and Investment Funds (CETOCOEN PLUS project: CZ.02.1.01/0.0/0.0/15_003/0000469 and the RECETOX research infrastructure: LM2015051 and CZ.02.1.01/0.0/0.0/16_013/0001761) and by the Ministry of Health, the Czech Republic (FNBr, 65269705).

Author Contributions

P.V. and E.B. designed the study, drafted the manuscript and interpreted the results; Kr.S. and L.M. processed the samples in the laboratory; B.Z. and Ka.S. performed bioinformatic analysis; E.B. and V.P. performed statistical data analysis. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49520-3.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[*17*] Zwinsová B, Petrov VA, Hrivňáková M, Smatana S, Micenková L, Kazdová N, Popovici V, Hrstka R, Šefr R, Bencsiková B, Zdražilová-Dubská L, Brychtová V, Nenutil R, Vídeňská P, **Budinská E.** Colorectal Tumour Mucosa Microbiome Is Enriched in Oral Pathogens and Defines Three Subtypes That Correlate with Markers of Tumour Progression. Cancers (Basel). 2021 Sep 25;13(19):4799. doi: 10.3390/cancers13194799. PMID: 34638284; PMCID: PMC8507728.

cancers

MDPI

Article

Colorectal Tumour Mucosa Microbiome Is Enriched in Oral Pathogens and Defines Three Subtypes That Correlate with Markers of Tumour Progression

Barbora Zwinsová [1,2,3], Vyacheslav A. Petrov [2], Martina Hrivňáková [1,2], Stanislav Smatana [2,4], Lenka Micenková [2], Natálie Kazdová [2], Vlad Popovici [2], Roman Hrstka [1], Roman Šefr [1], Beatrix Bencsiková [1], Lenka Zdražilová-Dubská [5,6], Veronika Brychtová [1], Rudolf Nenutil [1], Petra Vídeňská [1,2] and Eva Budinská [1,2,*]

[1] Research Centre for Applied Molecular Oncology (RECAMO), Masaryk Memorial Cancer Institute, 656 53 Brno, Czech Republic; zwinsova@recetox.muni.cz (B.Z.); martina.hrivnakova@mou.cz (M.H.); hrstka@mou.cz (R.H.); sefr@mou.cz (R.Š.); bencsikova@mou.cz (B.B.); vebrychtova@mou.cz (V.B.); nenutil@mou.cz (R.N.); petra.videnska@recetox.muni.cz (P.V.)

[2] Research Centre for Toxic Compounds in the Environment (RECETOX), Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic; viacheslav.petrov@recetox.muni.cz (V.A.P.); ismatana@mail.muni.cz (S.S.); lenka.micenkova@recetox.muni.cz (L.M.); mnau@mail.muni.cz (N.K.); vlad.popovici@recetox.muni.cz (V.P.)

[3] Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, 625 00 Brno, Czech Republic

[4] Research Centre of Information Technology, IT4Innovations Centre of Excellence, Brno University of Technology, 601 90 Brno, Czech Republic

[5] Department of Pharmacology, Faculty of Medicine, Masaryk University, 625 00 Brno, Czech Republic; lzd@mail.muni.cz

[6] Department of Laboratory Medicine-Clinical Microbiology and Immunology, University Hospital Brno, 625 00 Brno, Czech Republic

[*] Correspondence: eva.budinska@mou.cz

check for updates

Citation: Zwinsová, B.; Petrov, V.A.; Hrivňáková, M.; Smatana, S.; Micenková, L.; Kazdová, N.; Popovici, V.; Hrstka, R.; Šefr, R.; Bencsiková, B.; et al. Colorectal Tumour Mucosa Microbiome Is Enriched in Oral Pathogens and Defines Three Subtypes That Correlate with Markers of Tumour Progression. *Cancers* **2021**, *13*, 4799. https://doi.org/10.3390/cancers13194799

Academic Editors: Paolo Nuciforo and Emad M. El-Omar

Received: 27 August 2021
Accepted: 20 September 2021
Published: 25 September 2021

Simple Summary: Dysbiosis of the gut microbiome may contribute to the heterogeneity of colorectal cancer from phenotypic, prognostic and response to treatment perspectives. We analysed CRC microbiome by 16S rRNA gene sequencing of paired tumour mucosa, adjacent visually normal mucosa and stool swabs of 178 patients with stage 0–IV CRC. We observed that tumour mucosa is dominated by pathogenic bacteria of oral origin and proposed a CRC tumour microbiome subtyping system. The subtypes and tumour mucosa genera were associated with prognostic clinical covariates (tumour grade, localisation, TNM, *BRAF* mutation and MSI). In contrast, changes in the stool microbiome were associated with lymph node involvement and the presence of synchronous metastases. We discovered new associations between microorganisms and CRC and clinical parameters. Our study represents a step forward in understanding the role of the microbiome and its interactions with factors involved in tumour progression, and it opens novel avenues for exploring new treatments and biomarkers.

Abstract: Long-term dysbiosis of the gut microbiome has a significant impact on colorectal cancer (CRC) progression and explains part of the observed heterogeneity of the disease. Even though the shifts in gut microbiome in the normal-adenoma-carcinoma sequence were described, the landscape of the microbiome within CRC and its associations with clinical variables remain under-explored. We performed 16S rRNA gene sequencing of paired tumour tissue, adjacent visually normal mucosa and stool swabs of 178 patients with stage 0–IV CRC to describe the tumour microbiome and its association with clinical variables. We identified new genera associated either with CRC tumour mucosa or CRC in general. The tumour mucosa was dominated by genera belonging to oral pathogens. Based on the tumour microbiome, we stratified CRC patients into three subtypes, significantly associated with prognostic factors such as tumour grade, sidedness and TNM staging, *BRAF* mutation and MSI status. We found that the CRC microbiome is strongly correlated with the grade, location and stage, but these associations are dependent on the microbial environment. Our study opens new research avenues in the microbiome CRC biomarker detection of disease progression while identifying its limitations, suggesting the need for combining several sampling sites (e.g., stool and tumour swabs).

Keywords: colorectal cancer; 16S rRNA gene; tumour microbiome; microbial subtypes

1. Introduction

Colorectal cancer (CRC) is the third most frequent cancer worldwide and the second leading cause of cancer mortality in Europe [1]. It is a heterogeneous disease, both from a phenotypic and a prognostic and response to treatment perspective. The current standard treatments are limited and remain ineffective for many CRC patients due to inadequate patient selection, resulting in unneeded toxicity and high cost resulting from over-treating of patients that do not benefit [2,3]. Recent research shows that gut microbiota may significantly influence colorectal tumour initiation and progression [4–22].

Several studies showed that bacteria adherent to colorectal adenomas or carcinomas were different from bacteria adherent to healthy gut mucosa [8,11,12] due to the altered tumour environment with decreased pH and modified metabolic conditions resulting from hypoxia and onset of necrosis [23]. Gut microbiota can promote colon cancer development or change the tumour invasion potential through (i) immunomodulation [10,24–26] or (ii) metabolic activity—via the production of specific toxins inducing DNA damage responses. Overall, the evidence of microbiome importance in colon cancer development is so overwhelming that a bacterial driver-passenger model for colorectal cancer development and progression has been suggested [27] as an alternative to the universally accepted driver-passenger mutational adenoma-carcinoma model. Additionally, gut microbiota seems to play a crucial role also in response to anti-cancer therapy [28].

Previous studies associating gut dysbiosis with CRC were focused on comparing the gut microbiome in the normal-adenoma-carcinoma sequence [4–22,29–32]. It is the landscape of the microbiome within the ongoing disease and its associations with clinical variables that remain under-explored. The published studies vary in techniques employed, specimen origin and sample size, thus hampering any integrative analysis. Most studies compared diseased and healthy subjects, and the few that tried to characterise microbial composition within the CRC patients suffered from a small sample size. The specimens used in most studies were stool [4,6,7,15,17,18,20–22] or mucosa samples from colonoscopy biopsies [11,13,15] or post-resection [6,12,16,19]. Stool microbiota sampling has the advantage of being non-invasive, allowing its use for screening and follow-up studies. Some efforts combined information about the tumour-associated microbiome with existing prognostic scores in an attempt to improve the prediction accuracy [18] or to develop a new screening/prognostic model [33]. The results of two different meta-analyses showed that the accuracy of predicting diseased state was about 0.8, such as occult blood test results, the main non-invasive clinical test for this type of cancer [34,35]. However, the microbial composition in stool only partially reflects the situation in tumour mucosa, a trend consistent across different nationalities of the patients, sampling techniques or sequencing methodology [36].

The microbiota adherent to the mucosal tissue differs from the faecal microbiota in its needs for oxygen and nutrient types [37,38]. Therefore, the information derived from stool may be insufficient for capturing tumour-microbe interactions consistent with the disease prognosis. The relevance of the tumour mucosa microbiome assessment for screening purposes is dependent not only on the co-presence of the bacteria in both tumour mucosa and stool but also on its association with relevant clinical parameters in both sample types. Additionally, studying the (dis)similarity of bacterial composition between tumour and visually normal mucosa from the same individual may provide hints regarding the changes in microenvironment which have occurred favouring the growth of certain species and shed some light on the underlying tumour-immune system-microbe interactions and metabolic pathways.

Recently, two studies provided a comparison of bacterial composition in both tumour tissue and visually normal tissue and the bacterial composition of stool samples from the same patients [34,35]. Liu et al. [34] showed that the bacterial communities in both tumour tissue and visually normal tissue were similar. Still, the study was vastly underpowered (n = 8 individuals) and did not explore the clinical relevance of this similarity. Other studies associated microbiome on tumour or in stool with clinical variables [35,39,40] but had a similar disadvantage in terms of statistical power (n = 25, n = 30, n = 53, individuals, respectively).

The studies mentioned were species-centric because they compared the abundance of individual microbial species between the groups of interest. However, a broader view is needed to account for lesser-known species coupled with a larger sample size allowing for capturing enough inter-tumour heterogeneity, thus better understanding the possible effects of bacteria on tumour growth, aggressiveness or response to therapy. Our study takes a microbial community-centric approach to provide a comprehensive description of the CRC tumour microbiome based on 16S rRNA sequencing. We analyse three sample types (tumour mucosa, visually normal mucosa, stool) from n = 178 individuals with stage 0–IV colorectal cancer.

Our study has a dual nature, both exploratory and confirmatory. We explore and interpret the landscape of the tumour mucosa-associated microbiome with respect to clinical variables and microbial composition of paired adjacent visually normal mucosa and paired stool samples. Benefitting from a larger sample size, we advance the state-of-the-art knowledge by reporting previously unseen associations. Most importantly, we capture the tumour microbial heterogeneity and derive CRC tumour microbiome subtypes.

2. Materials and Methods

2.1. Patients and Specimens

All specimens were collected at Masaryk Memorial Cancer Institute (Brno, Czech Republic) from 2015 to 2019. Patient inclusion criteria were (i) scheduled for resection based on preliminary screening (such as a colonoscopy), (ii) no neoadjuvant treatment, (iii) no previous CRC diagnosis (iv) with confirmed stage 0–IV CRC without multiplicities (single tumour). The stool samples were collected from untreated patients before the scheduled surgery. Patients performed the collection at home, the morning of their hospitalisation for the surgery and brought the samples to the hospital, where they were immediately frozen at −80 °C until further processing. Swabs from the tumour and visually normal mucosa were collected within 30 minutes of the tumour resection at the pathology department. Whenever possible, the swab from visually normal tissue was taken at least 20 cm proximally to the tumour. The swabs were then stored immediately in a freezer at −20 °C and, without unnecessary delay, transferred to −80 °C until further processing. All samples, including stool, were collected using DNA free cotton swabs (Deltalab, Barcelona, Spain).

Overall, we analysed n = 483 samples from n = 178 CRC patients. There were 127 triplets (all three sample types from the same patient) and 51 mucosa duplets (swabs from tumour and visually normal mucosa from the same patient).

The study was approved by the ethical committee of Masaryk Memorial Cancer Institute. All patients gave written informed consent following the Declaration of Helsinki prior to participating in the study.

2.2. DNA Extraction, PCR Amplification and Sequencing of 16S rRNA Gene

According to the manufacturer's instructions, the DNA extraction was performed using DNeasy® PowerSoil® Isolation kit (QIAGEN, Düsseldorf, Germany). Extracted DNA was used as a template in amplicon PCR to target the V4 hypervariable region of the bacterial 16S rRNA gene. The 16S metagenomics library was prepared according to the 16S Metagenomic Sequencing Library Preparation protocol (Illumina, San Diego, CA, USA), with some deviations described below. Each PCR was performed with HotStarTaq Master Mix Kit (QIAGEN, Hilden, Germany) in triplicate, with the primer pair consisting of Illumina overhang nucleotide sequences, an inner tag, and gene-specific sequences [41,42]. The Illumina overhang served to ligate the Illumina index and adapter. Each inner tag, i.e., a unique sequence of 7–9 bp, was designed to differentiate samples into groups. Primer sequences and PCR cycling conditions are summarised in Table S3. After PCR amplification, triplicates were pooled, and the amplified PCR products were determined by gel electrophoresis. PCR clean-up was performed with Agencourt AMPure XP beads (Beckman Coulter Genomics, Danvers, MA, USA). Samples with different inner tags were equimolarly pooled based on fluorometrically measured concentration using Qubit® dsDNA HS Assay Kit (Invitrogen™, Carlsbad, CA, USA) and microplate reader (Synergy Mx, BioTek, Winooski, VT, USA). Pools were used as a template for a second PCR with Nextera XT indexes (Illumina, USA). Differently indexed samples were quantified using the qPCR kit KAPA Library Quantification Complete Kit (Roche, Indianapolis-Marion County, IN, USA) and LightCycler 480 Instrument (Roche, USA) and equimolarly pooled according to the measured concentration. The prepared libraries were checked with a 2100 Bioanalyzer Instrument using the High Sensitivity D5000 Screen tape (Agilent Technologies, Santa Clara, CA, USA), and concentration was measured with qPCR shortly prior to sequencing. The final library was diluted to a concentration of 8 pM, and 20% of PhiX DNA (Illumina, USA) was added. According to the manufacturer's instructions, sequencing was performed with the Miseq reagent kit V2 (500 cycles) using a MiSeq instrument (Illumina, USA).

2.3. Data Analysis

2.3.1. Preprocessing and Quality Control

Forward and reverse pair-end reads were demultiplexed, and barcodes and primers were trimmed. Denoising algorithm with DADA2 [43] was applied separately on forward and reverse reads that passed the quality and length filter and did not contain N's. Reads were merged using the fastq-join method [44]. In the next step, chimaeras were detected with the function removeBimeraDenovo in DADA2. Chimaera sequences were subsequently excluded from the analysis, and Amplicon Sequence Variant (ASV) table was created.

After quality filtering and chimaeras removing, the number of reads ranged from 2968 to 239,116, with a median of 44,371 and a mean of 53,074 reads per sample. The number of reads did not differ between the sample types (paired Wilcoxon test, Figure S1).

2.3.2. Taxonomy Assignment and Metabolic Potential Prediction

Taxonomy was assigned to each ASV based on SILVA 123 reference database [45] using the algorithm UCLUST [46] in QIIME [47]. BLAST algorithm [48] was used to identify the species, and all taxa with the maximum identity and minimum *e*-value were selected for each ASV. The observed species metric and the Chao1 and Shannon index were used to estimate alpha diversity for each sample in QIIME. Beta diversity was computed in QIIME using both weighted and unweighted UniFrac metrics [49].

We filtered out the ASVs unassigned at the phylum level and all the ASVs belonging to the phylum Cyanobacteria. Only the taxa present in at least three samples of the same sample type and at the same time represented by at least nine reads were kept for further analysis to account for possible contaminations. The threshold of 9 reads represents 0.3% taxa abundance in the sample with the least number of reads (2968).

This filtering step discarded 46–55% of taxa at each taxa level (Table S4).

Picrust 2 [50] was used to predict hypothetical abundances of *KEGG* orthologs in each sample and to summarise them into higher functional processes.

2.3.3. Statistical Analysis and Data Mining

All comparisons between the three sample types were performed on triplet samples from 127 patients, totalling $n = 381$ samples for the analysis. For the analysis of tumour-visually normal mucosa pairs, we used paired tumour and visually normal mucosa swabs from 178 patients (totalling 356 samples). We used all the available samples for analyses performed within each sample type (178 for tumour mucosa swabs, $n = 178$ for visually normal tissue mucosa swabs and $n = 127$ for stool).

Data were analysed using appropriate corrections and approaches for compositional data [51–54]. Zero multiplicative replacement [53] was applied prior to the centred log-ratio (clr) transformation.

Non-metric Multidimensional Scaling (R vegan package [55]) over Aitchinson distance matrices (R coda.base package [56]) was used to analyse tumour microbial heterogeneity and β-diversity. To estimate the contribution of clinical traits in the microbiome, β-diversity permutational multivariate analysis of variance for distance matrices (R adonis function of vegan 2.5.4 package [55]) with 999 permutations were used. To assess the differences between the sample types in alpha diversity, we used a paired non-parametric two-way Mann-Whitney U test. We applied a non-parametric approach to identify differences in microbial composition between sample types and the associations between relative microbial abundance and clinical variables. For non-parametric analysis, the Friedman test with paired Wilcoxon test and rank regression was used (R package Rfit [57]). A drop in dispersion test was used to produce overall p-values for rank regression models. The Cochran Q test was used to analyse differences in the presence of genera across sample types (analysis of triplets). Benjamini-Hochberg correction for multiple hypothesis testing was applied [58]. Results were considered significant at FDR <0.1. The adjusted p-values are referred to as q-values. Visualisation was performed with gplots 3.0.1.1, ggplot2, ComplexHeatmap 1.17.and circlize 0.4.8 packages [59–62].

For each clinical variable (or a combination thereof), we only tested genera present in at least 10 samples in one clinical group (or a combination thereof). We do emphasise that we approached this statistical testing from the point of view of a pilot discovery study.

Due to the known association between tumour grade and location [63] (also confirmed in our data, $p < 0.001$, Fisher's exact test), we investigated the associations of the microbiome with grade and tumour location in a model with the interaction between covariates compared to a model without interaction. To ensure a more balanced design, we considered three locations: right and transverse, left, rectosigmoid and rectum, respectively.

The threshold of false discovery rate was set to 0.1, as is customary in similar studies, with the aim to identify potential candidates for further research. While we consider only associations with FDR <0.1 to be statistically significant, we also report the unadjusted results $p < 0.05$ for hypothesis confirmation by other studies.

2.4. Data Access

The data were uploaded to the European Nucleotide Archive under accession number PRJEB35990.

2.5. Validation

We performed partial validation of our results on three publicly available datasets. The association of the tumour microbiome with tumour localisation was validated in the dataset of Dejea et al. [31], $n = 23$. No grade information was available, and hence in the validation we did not use the grade*localisation interaction term. Publicly available fastq files were analysed with QIIME pipeline with the appropriate approach for 454 Roche sequencing. Taxonomy was assigned using SILVA 123 database to have comparable results with our dataset.

The association of stool microbiome with AJCC staging and TNM staging was validated in two datasets (Zeller et al. [32] and Feng et al. [30]). The processed datasets with taxonomic information were used as available in R package curatedMetagenomicData [64] and were normalised using the clr transformation before the analysis.

All associations were tested using rank regression (R package Rfit [57]). The dataset of Feng et al. only contained one M1 sample; hence we only analysed associations with AJCC staging, T stage and N stage.

3. Results

In our effort to describe tumour microbial landscape, we explored the differences in microbiome abundance, diversity, the presence/absence of the species and the proportion of samples with the respective genera in different sample types across patient groups defined by clinical variables (Table 1).

Table 1. Table of clinical variables and their distribution in the complete set of 178 patients, the subset of 127 patients and in the CRC tumour microbial subtypes, respectively. (For categorical variable, Fisher exact test was performed and for continuous data, Kruskal-Wallis test was used.).

Clinical Variables	Data Subset Comparison			Tumour Microbiome Subtypes			
	All Tumours (*n* = 178)	Triplets (*n* = 127)	*p*-Value	TMS1 (*n* = 46)	TMS2 (*n* = 55)	TMS3 (*n* = 77)	*p*-Value
age at diagnosis	Mean (SD)	Mean (SD)	0.804	Mean (SD)	Mean (SD)	Mean (SD)	0.887
	66.92 (10.66)	66.61 (10.61)	-	66.89 (9.88)	67.47 (11.39)	66.55 (10.69)	-
gender	*n* (%)	*n* (%)	1	*n* (%)	*n* (%)	*n* (%)	0.729
male	99 (55.6)	70 (55.1)	-	25 (54.3)	33 (60.0)	41 (53.2)	-
female	79 (44.4)	57 (44.9)	-	21 (45.7)	22 (40.0)	36 (46.8)	-
tumour localisation	*n* (%)	*n* (%)	0.597	*n* (%)	*n* (%)	*n* (%)	<0.001
right	64 (36.0)	48 (37.8)	-	28 (60.9)	11 (20.0)	25 (32.5)	-
transverse	19 (10.7)	13 (10.2)	-	6 (13.0)	5 (9.1)	8 (10.4)	-
left	44 (24.7)	36 (28.3)	-	4 (8.7)	17 (30.9)	23 (29.9)	-
rectosigmoideum	32 (18.0)	23 (18.1)	-	6 (13.0)	10 (18.2)	16 (20.8)	-
rectum	19 (10.7)	7 (5.5)	-	2 (4.3)	12 (21.8)	5 (6.5)	-
grade	*n* (%)	*n* (%)	0.998	*n* (%)	*n* (%)	*n* (%)	<0.001
NA, in situ	7 (3.9)	5 (3.9)	-	0 (0.0)	3 (5.5)	4 (5.2)	-
1	18 (10.1)	12 (9.4)	-	1 (2.2)	5 (9.1)	12 (15.6)	-
2	102 (57.3)	73 (57.5)	-	18 (39.1)	37 (67.3)	47 (61.0)	-
3	51 (28.7)	37 (29.1)	-	27 (58.7)	10 (18.2)	14 (18.2)	-
AJCC stage	*n* (%)	*n* (%)	0.968	*n* (%)	*n* (%)	*n* (%)	0.136
0	8 (4.5)	6 (4.7)	-	0 (0.0)	3 (5.5)	5 (6.5)	-
I	31 (17.4)	26 (20.5)	-	2 (4.3)	12 (21.8)	17 (22.1)	-
II	66 (37.1)	45 (35.4)	-	21 (45.7)	19 (34.5)	26 (33.8)	-
III	48 (27.0)	34 (26.8)	-	16 (34.8)	12 (21.8)	20 (26.0)	-
IV	25 (14.0)	16 (12.6)	-	7 (15.2)	9 (16.4)	9 (11.7)	-
tumour pathologic stage	*n* (%)	*n* (%)	0.979	*n* (%)	*n* (%)	*n* (%)	0.007
pTis	8 (4.5)	6 (4.7)	-	0 (0.0)	3 (5.5)	5 (6.5)	-
pT1	11 (6.2)	10 (7.9)	-	0 (0.0)	5 (9.1)	6 (7.8)	-
pT2	32 (18.0)	24 (18.9)	-	2 (4.3)	12 (21.8)	18 (23.4)	-
pT3	115 (64.6)	79 (62.2)	-	42 (91.3)	30 (54.5)	43 (55.8)	-
pT4	12 (6.7)	8 (6.3)	-	2 (4.3)	5 (9.1)	5 (6.5)	-
regional lymph nodes pathologic stage	*n* (%)	*n* (%)	0.618	*n* (%)	*n* (%)	*n* (%)	0.041
pN0	109 (61.2)	79 (62.2)	-	23 (50.0)	36 (65.5)	50 (64.9)	-
pN1	46 (25.8)	36 (28.3)	-	13 (28.3)	10 (18.2)	23 (29.9)	-
pN2	23 (12.9)	12 (9.4)	-	10 (21.7)	9 (16.4)	4 (5.2)	-
synchronous distant metastasis	*n* (%)	*n* (%)	0.846	*n* (%)	*n* (%)	*n* (%)	0.722
M0	153 (86.0)	111 (87.4)	-	39 (84.8)	46 (83.6)	68 (88.3)	-
M1	25 (14.0)	16 (12.6)	-	7 (15.2)	9 (16.4)	9 (11.7)	-

Table 1. *Cont.*

Clinical Variables	**Data Subset Comparison**			**Tumour Microbiome Subtypes**			
	All Tumours (*n* = 178)	**Triplets (*n* = 127)**	***p*-Value**	**TMS1 (*n* = 46)**	**TMS2 (*n* = 55)**	**TMS3 (*n* = 77)**	***p*-Value**
MSI/MSS	*n* (%)	*n* (%)	1	*n* (%)	*n* (%)	*n* (%)	<0.001
MSI	27 (15.2)	19 (15.0)	-	16 (34.8)	4 (7.3)	7 (9.1)	-
MSS	110 (61.8)	81 (63.8)	-	22 (47.8)	37 (67.3)	51 (66.2)	-
NA	41 (23.0)	27 (21.2)	-	8 (17.4)	14 (25.4)	19 (24.7)	-
BRAF	*n* (%)	*n* (%)	1	*n* (%)	*n* (%)	*n* (%)	0.022
BRAF wt	77 (43.3)	53 (41.7)	-	17 (37.0)	27 (49.1)	33 (42.9)	-
BRAF mut	12 (6.7)	9 (7.1)	-	7 (15.2)	1 (1.8)	4 (5.2)	-
NA	89 (50.0)	65 (51.2)	-	22 (47.8)	27 (49.1)	40 (51.9)	-
KRAS	*n* (%)	*n* (%)	1	*n* (%)	*n* (%)	*n* (%)	0.839
KRAS wt	24 (13.5)	17 (13.4)	-	7 (15.2)	8 (14.5)	9 (11.7)	-
KRAS mut	13 (7.3)	9 (7.1)	-	5 (10.9)	4 (7.3)	4 (5.2)	-
NA	141 (79.2)	101 (79.5)	-	34 (73.9)	43 (78.2)	64 (83.1)	-
NRAS	*n* (%)	*n* (%)	1	*n* (%)	*n* (%)	*n* (%)	0.553
NRAS wt	37 (20.8)	26 (20.5)	-	11 (23.9)	12 (21.8)	14 (18.2)	-
NRAS mut	2 (1.1)	1 (0.8)	-	1 (2.2)	1 (1.8)	0 (0.0)	-
NA	139 (78.1)	100 (78.7)	-	34 (73.9)	42 (76.4)	63 (81.8)	-

CRC—colorectal cancer, TMS—tumour microbial subtypes, SD—standard deviation, NA—not available, pT—tumour pathologic stage, pTis—tumour in situ, pN—regional lymph nodes pathologic stage, M—synchronous distant metastasis, MSI—microsatellite instability MSS—microsatellite stable, wt—wild type, mut—mutation.

3.1. Microbial Categorisation According to Sample Type

There was no significant difference between the read counts across different sample types (paired analysis of sample triplets, see Methods).

The analysis of the 127 triplet samples revealed that the microbial diversity was significantly decreased in mucosal samples (both tumour mucosa and visually normal mucosa swabs) compared to stool, as measured by the number of observed species, Chao 1 and Shannon index (Figure S1). No differences were found between the tumour mucosa swabs and visually normal mucosa swabs.

Overall, in all the 483 samples, we identified 5449 ASVs: of these, 4800 ASVs in the 127 triplet samples. The QIIME assigned species only to 48 ASVs. Hence, we also performed a manual BLAST search to the SILVA database (Table S5).

For further analysis, however, we operated on higher taxonomic levels. After the taxa filtering step (Table S4), 13 phyla, 25 classes, 43 orders, 75 families and 264 genera were identified in the 127 triplets, most of which in all three sample types (Table S6). Inclusion of the additional 51 duplets (tumour mucosa and visually normal mucosa swabs) resulted only in slight differences at the genus level—the identified taxa remained the same. What changed was their unique presence in some sample types (Text S1).

While most of the genera were found in all three sample types, their incidence and abundance across sample types varied greatly between mucosal samples and stool, both in overall and pairwise comparisons (Text S1, Figure S16). In this case, 14 genera (*Stomatobaculum, Pseudoramibacter, Pelomonas, Pasteurella, Mycoplasma, Kingella, Johnsonella, Helicobacter, Deinococcus, Centipeda, Bergeyella, Actinobacillus, Abiotrophia* and an unassigned genus from order *Comamonadaceae*) were detected only in mucosal (tumour and visually normal) samples (Figure S2).

We further analysed the pairwise incidence of the 264 genera across sample types. We found that 104 genera varied significantly across sample types (analysis of 127 triplets, Text S1, Table S7).

To categorise the microbial genera based on their preferred environment: we compared their abundance across sample types. Of the 264 genera, 121 differed significantly in abundance across the sample types (Tables S8 and S9, Figure S3). Based on these results, we defined five microbial categories (Figure 1). The first is based solely on the results of tumour vs stool comparison: tumour genera (57 genera, more abundant in tumours than stool). Additionally, within the category of tumour genera, we defined mucosa genera (52 genera, also enriched in visually normal mucosa compared to stool) and tumour-specific genera (16 genera of tumour category, additionally enriched in tumours compared to visually normal mucosa). In this case, 49 genera were significantly more abundant in stool than tumours and visually normal mucosa from the group of stool genera. The fifth category was defined as the no-difference genera (143 genera, no difference across any of the sample types) (Text S1).

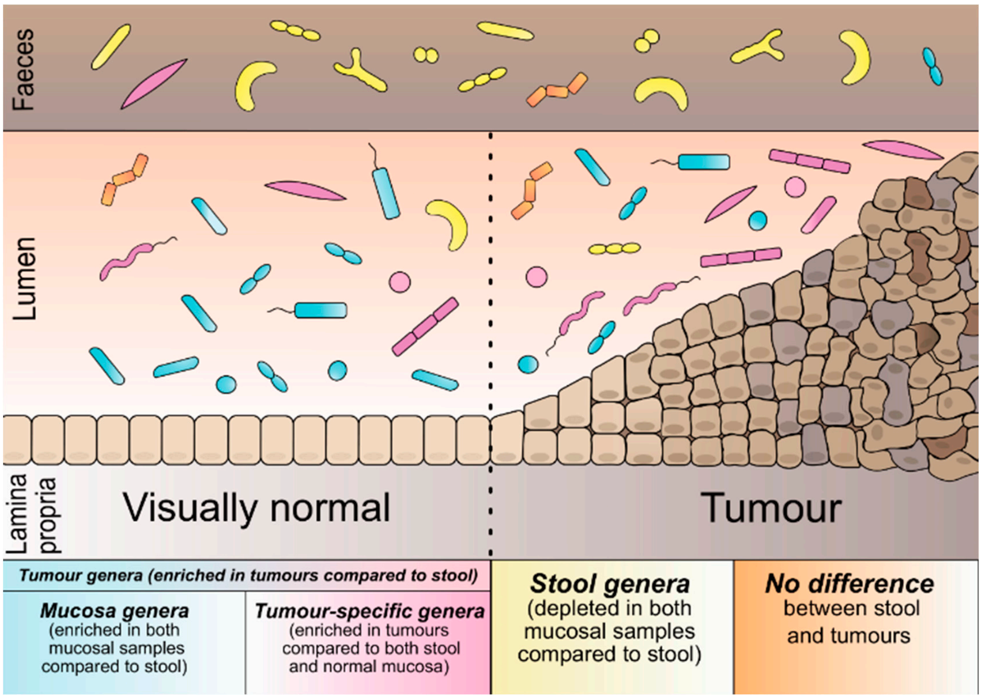

Faeces
Lumen
Lamina propria
Visually normal
Tumour
Tumour genera (enriched in tumours compared to stool)
Mucosa genera (enriched in both mucosal samples compared to stool)
Tumour-specific genera (enriched in tumours compared to both stool and normal mucosa)
Stool genera (depleted in both mucosal samples compared to stool)
No difference between stool and tumours

Figure 1. Schematic representation of bacterial categories according to their preferred environment.

3.2. The Landscape of CRC Tumour Microbiome

For the description of tumour mucosa microbial heterogeneity without stool contaminants, we only considered species that were statistically significantly enriched in tumour mucosa compared to stool. We hence investigated the group of 57 tumour genera with a special focus on the subgroup of 16 tumour-specific genera (*Gemella, Granulicatella, Parvimonas, Hungatella, Peptoclostridium, Flavonifractor, Selenomonas 3, Fusobacterium, Leptotrichia, Eikenella, Campylobacter, Slackia, Streptococcus, Howardella, Solobacterium, Defluviitaleaceae UCG-011,* Figure 2A).

We performed the analysis of co-occurrence and observed significantly increased co-occurences between 20 tumour genera (of which 13 tumour-specific) (Text S1, Figure S4, Table S10). We also observed 14 significantly decreased co-occurrences between genera (Text S1, Figure S4, Table S10).

Tumour genera incidence ranged from 1.1% to 99.4% (median 26.4%) of tumours with the median abundance of the individual genera in the samples with the genus detected ranging from 0.01% to 29.8% (median 0.15%) (Figure 2A). Overall, tumour genera constituted 1.1% to 97% (median 59.6%) while the tumour-specific genera constituted between 0.0–62.3% (median 3.1%) of the microbiome found on tumour mucosa (Figure 2C).

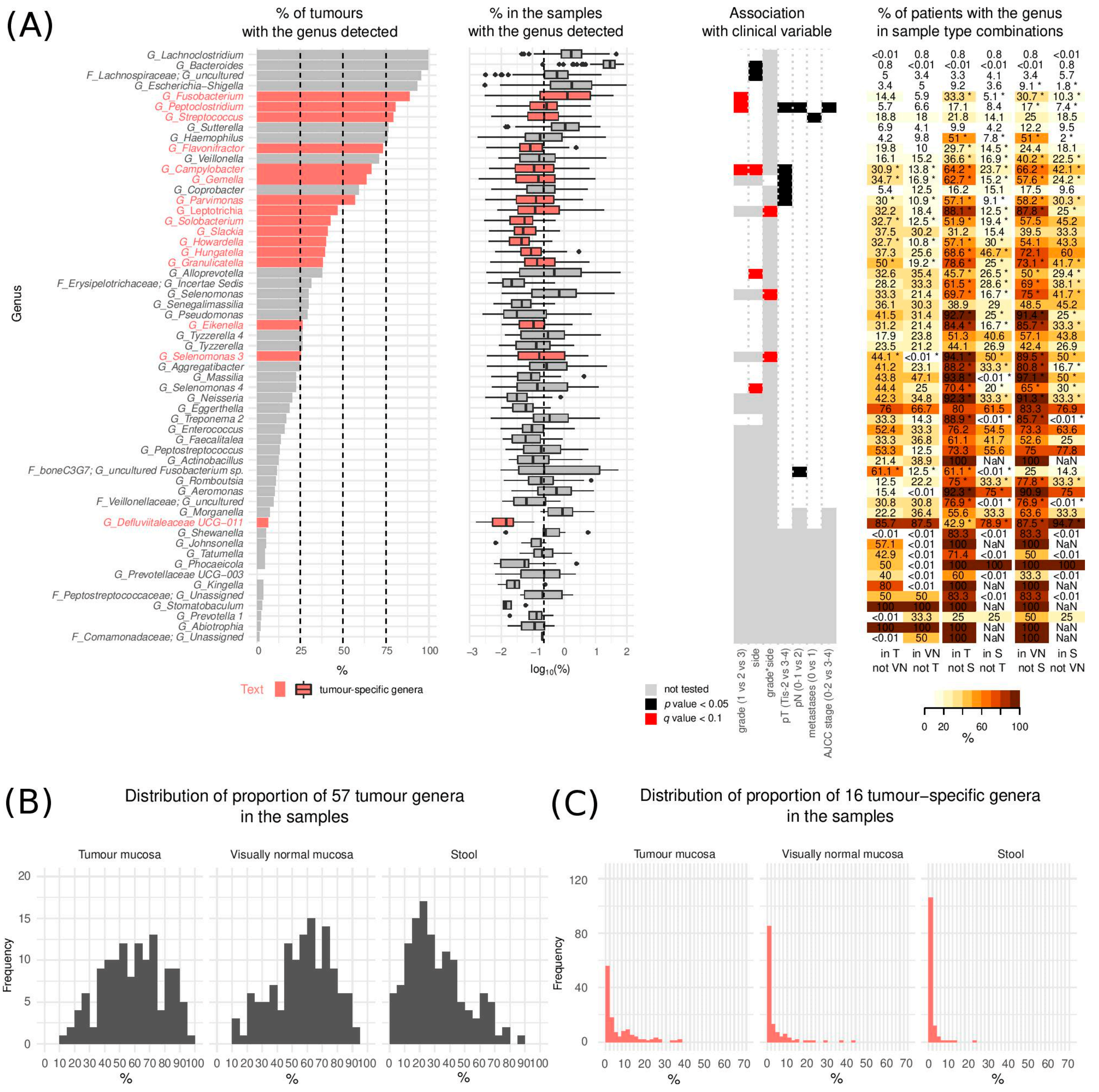


Figure 2. Tumour genera. (**A**) (left) Proportion of tumours with the genera detected and distribution of their respective relative abundances (in %) in the samples where they were detected (middle) and association of the bacteria with clinical variables (right). The vertical dashed line represents the median relative abundance of all 264 detected genera (median = 0.24%). The boxplot middle vertical line represents median, the box represents the interquartile range (IQR), the whiskers extend to +/−1.5 IQR. The black dots refer to outliers. (**B**) Overall proportion of the 57 tumour genera in the three sample types (*n* = 127) (**C**) Overall proportion of the 16 tumour-specific genera in the three sample types (*n* = 127). (Tis—Tumour in situ, pT—tumour pathologic stage, pN—regional lymph nodes pathologic, T—tumour swabs, VN—visually normal mucosa swabs, S—stool, NaN—not a number). * $p < 0.05$.

We performed a detailed literature search (Table S11) which revealed that tumour genera consisted predominantly of oral bacteria, many known as oral pathogens.

Some of the tumour genera of (possible) oral origin identified in our study, while previously associated with CRC, were never reported on tumour mucosa, namely *Solobacterium* (increased in CRC faecal samples [35]), *Slackia* and *Pseudomonas* (decreased [12,19] in CRC faecal samples), and *Treponema* (the presence of which in the oral cavity was associated with increased risk of CRC [33]).

We newly identified many genera of both oral and gut origin, not previously associated with CRC, with increased abundance in the tumour lesions: *Selenomonas 3, Selenomonas 4, Aggregatibacter, Actinobacillus, Bergeyella, Phocaeiola, Defluviitaleaceae UCG-011, Abiotrophia, Johnsonella, Stomatobaculum, Kingella, Shewanella, Tatumella, Senegalimassilia, Aeromonas, Prevotellaceae UCG-003, Incertae Sedis* genus from family *Erysipelotrichaceae,* an uncultured species from *Veillonelaceae* family, and an uncultured species from *bone*C3G7 at the family level (BLAST hit *Fusobacterium necrophorum*) (oral origin) and *Tyzzerella 4, Massilia* and an unassigned genus from *Peptostreptococcaceae* family (gut origin).

Amongst tumour genera of gut origin, *Lachnoclostridium*, *Flavonifractor* [65,66], *Sutterella* and *Hungatella* (ex-*Clostridium hathewayi*) [67] were previously only reported increased in the stool of patients with CRC.

3.3. Microbiome and Clinical Variables

Prior to the subtype derivation, we assessed the association of bacterial genera from all the sampled environments with the clinical parameters and interpreted the results based on our microbial categorisation. We performed partial validation on three publicly available datasets.

β-diversity analysis by NMDS performed on each sample type separately showed that tumour location was the factor with the highest influence on total microbiome composition for all sample types, while tumour histological grade affected only tumour samples (Text S1, Table S12, Figures S5 and S6).

The results of regression analysis for each clinical variable are summarised in Table 2 and Table S13, and Figures S7–S10; the detailed results of partial validation are presented in Text S2, Tables S1, S2 and S14.

Increased abundance of *Fusobacterium, Campylobacter* and *Leptotrichia* in tumour mucosa appeared to be independent predictors of tumour's higher grade ($p < 0.01$, FDR < 0.1). *Leptotrichia* was significantly increased on visually normal mucosa adjacent to grade 3 left-sided tumours ($p < 0.05$, FDR < 0.1).

The mucosa of grade 3 right-sided tumours was enriched in *Prevotella, Selenomonas* and *Selenomonas 3* ($p < 0.01$, FDR < 0.1). *Prevotella* was also increased in the stool of patients with grade 3 rectosigmoid/rectum tumours ($p < 0.01$, FDR < 0.1). The mucosa of grade 3 tumours of the rectosigmoid/rectum and visually normal mucosa adjacent to left, rectosigmoid and rectal tumours, regardless of the grade, were enriched in *Lachnospira* ($p < 0.05$, FDR < 0.1).

The mucosa of left-sided (for some including rectosigmoid/rectum) low-grade tumours was enriched in *Ruminiclostridium 6, Coprococcus 2*, [*Eubacterium] ventriosum* group, *Clostridiales Vadin BB60* group, *Ruminococcaceae UCG-010* and an uncultured species and an *Incertae Sedis* genus from the *Lachnospiraceae* family ($p < 0.01$, FDR < 0.1). *Ruminoclostridium 6* remained enriched also in the stool of patients with grade 2 left-sided, rectosigmoid and rectal tumours ($p < 0.01$, FDR < 0.1). *Methanobrevibacter*, *Victivallis* were significantly enriched in the mucosa of low-grade tumours of rectosigmoid and rectum (both $p < 0.01$, FDR < 0.1).

Christensenellaceae R-7 group, *Bifidobacterium* and *Ruminococcaceae UCG-013* were increased in mucosa of the left-sided, rectosigmoid and rectal tumours ($p < 0.01$, FDR < 0.1). Similar associations were found for visually normal mucosa for *Christensenellaceae R-7 group*, *Coprococcus 1*, *Lachnospira* and *Bifidobacterium* ($p < 0.01$, FDR < 0.1). The increased abundance of the *Christensenellaceae R-7* group in tumour mucosa of left-sided tumours was also validated in an independent dataset ($p = 0.0047$) (Text S2, Table S14). When comparing early (0–II) and advanced (III–IV) stages, we identified an increased abundance of *Akkermansia* in the stool of advanced stage tumours ($p < 0.01$, FDR < 0.1) (Table S13, Figures S11 and S12).

Table 2. Summary of rank regression results ($p < 0.05$) associating microbiome of the three different sample types with the clinical variables. **Bold text** denotes genera significant at FDR < 0.1, text underlined by a solid line denotes that the association was validated in an independent dataset, marked by the superscript number (^{1} Feng et al. [30]; ^{2} Dejea et al. [31], * previously published association [40,66,68–72], see Discussion and Table S11). Up and down arrows denote increase or decrease in abundance, respectively.

Regression Covariate	Effect/Contrast	Tumour Mucosa	Visually Normal Mucosa	Stool
grade	increasing grade	↑ ***Fusobacterium*** *, ***Campylobacter*** *, ***Leptotrichia***, *Peptoclostridium*, *Mogibacterium* *	-	-
		-	↓ Unassigned genus from order *Opitutae vadin HA64*	-
location	right-sided/transverse vs left-sided and rectum/rectosigmoid	↑ *Holdemania, Selenomonas 4, Clostridium sensu stricto 1, Alloprevotella*	↑ ***Selenomonas 3, Selenomonas, Treponema 2***	-
		↓ ***Bifidobacterium*** *, ***Christensenellaceae R-7*** **group** ^{2}, ***Ruminococcaceae UCG-013***, *Fusicatenibacter*	↓ ***Lachnospira, Bifidobacterium, Coprococcus 1, Christensenellaceae R-7*** **group**	-
	right-sided/transverse vs left-sided	↑ *Campylobacter*, *Alloprevotella*	-	-
		↓ Family *XIII AD3011* group, *Coprococcus 1*	-	
	right-sided/transverse vs rectosigmoid/rectum	↑ *Oribacterium, Fretibacterium*	-	-
		-	↓ **[*Eubacterium*] *ventriosum* group**	
grade*location interaction	low-graded; right-sided/transverse	↑***Ruminococcaceae UCG-010*, uncultured bacterium from *Clostridiales vadinBB60* group**	-	↑ **Unassigned genus from order *Opitutae vadin HA64*, *Porphyromonas***
	grade 2; left-sided	↓***Coprococcus 2, Ruminiclostridium 6,*** **[*Eubacterium*] *ventriosum* group, *Incertae Sedis* from *Lachnospiraceae* family**	↓ ***Gemella, Corynebacterium 1***	↓ ***Ruminiclostridium 6***, *Coprococcus 2*
	grade 2; rectosigmoid/rectum	-	↑ ***Veillonella***	↑ ***Veillonella***
		↓ ***Methanobrevibacter, Dielma, Victivallis***	↓ ***Methanobrevibacter*, an uncultured genus from the *Peptococcaceae* family**	↓ ***Victivallis, Ruminiclosridium 6, Lachnospiraceae UCG-005*, an unassigned genus from order *Mollicutes* RF9**
	grade 3; right-sided/transverse	↑ ***Prevotella, Selenomonas, Selenomonas 3***	-	-

Table 2. *Cont.*

Regression Covariate	Effect/Contrast	Tumour Mucosa	Visually Normal Mucosa	Stool
grade*location interaction	grade 3; left-sided	-	↑ ***Eisenbergiella, Leptotrichia, Escherichia-Shigella, Veillonella***	↑ ***Veillonella, Prevotella 7***
		↓ ***Coprococcus 2, Ruminiclostridium 6, [Eubacterium] ventriosum* group, *Incertae Sedis* from *Lachnospiraceae* family, *Odoribacter***	↓ ***Gemella, Corynebacterium 1***	↓ ***Coprococcus 2***
	grade 3; rectosigmoid/rectum	↑ ***Lachnospira***	↑ ***Veillonella***	↑ ***Prevotella, Prevotella 7***
		↓ ***Methanobrevibacter, Dielma, Victivallis***	↓ ***Methanobrevibacter, Eisenbergiella*, an uncultured genus from the *Peptococcaceae* family**	↓ ***Lachnospiraceae UCG-005*, unassigned genus from order *Mollicutes RF9***
AJCC stage	III–IV vs 0–II	↑ *Peptoclostridium*	-	↑ ***Akkermansia***
		-	↓ *Gelria*	-
Tumour pathologic stage	pT 3–4 vs pTis-2	↑ ***Peptoclostridium, Gemella, Campylobacter, Parvimonas***	↑ ***Peptoclostridium, Escherichia-Shigella*, an unassigned species from *Ruminococcaceae***	↑ *Escherichia-Shigella*
		↓ ***Coprobacter***, *Intestinimonas, Ruminococcaceae UCG-009, Oscillospira, Cloacibacillus*	↓ ***Intestinimonas, Ruminococcaceae UCG-009, Holdemanella, Coprobacter***, *Gelria*, an uncultured genus from the *Christensenellaceae* family	↓ ***Prevotella 6***, *Ruminococcaceae UCG-011*[1]
Regional lymph nodes stage	N1–2 vs N0	↑ *Peptoclostridium*	-	↑ *Peptococcus, Campylobacter*, ***Akkermansia*** *, *Selenomonas, Porphyromonas* *, *Streptococcus, Oscillospira*
		↓ *Prevotellaceae UCG-001*, uncultured *Fusobacterium sp.* from family *boneC3G7*	↓ [*Eubacterium*] *hallii* group	↓ Faecalibacterium, Ruminiclostridium, Dorea *, Lachnospiraceae FCS020 group
Synchronous distant metastasis	present vs absent	↑ *Porphyromonas, Streptococcus, Ruminococcaceae UCG-005*	↑ *Akkermansia*	↑ **uncultured genus from *Erysipelotrichaceae* family**, *Akkermansia, Coprococcus 1, Solobacterium*
		-	↓ *Gelria*, [*Eubacterium*] *brachy* group, uncultured genera from *Christensenellaceae* family, *Gordonibacter, Fretibacterium*	↓ *Selenomonas, Ruminococcaceae UCG-004*

FDR-false discovery rate.

Patients with advanced T stages (pT 3–4) were characterised by a significant increase in abundance of *Gemella, Campylobacter, Peptoclostridium and Parvimonas* ($p < 0.01$, FDR < 0.1) on tumour mucosa, and increased *Peptoclostridium, Escherichia-Shigella* ($p < 0.01$, FDR < 0.1) in the adjacent visually normal mucosa. Early T stage tumours (pTis-2) were associated with an increase in *Coprobacter*, on tumour mucosa, increased *Intestinimonas*, *Ruminococcaceae UCG-009*, *Holdemanella* and *Coprobacter* on the adjacent visually normal tissue ($p < 0.05$, FDR < 0.1) and *Prevotella 6* ($p < 0.01$, FDR < 0.1) and *Ruminococcaceae UCG-011* ($p < 0.05$, FDR > 0.1) in the stool (Table S13, Figures S11 and S13). We validated the decrease of *Ruminococcaceae* in the stool of patients with pT 3–4 stage ($p = 0.00113$, Feng et al.) (Text S2, Table S14).

The presence of metastases (local or distant) at the time of diagnosis was predominantly associated with changes in the stool microbiome. Except for the increased abundance of *Akkermansia* in stool of patients with N1–2 stage tumours ($p < 0.01$, FDR < 0.1) and of the uncultured genus from the *Erysipelotrichaceae* family in the stool of patients with synchronous distant metastases ($p < 0.01$, FDR < 0.1), none of these associations were significant after FDR correction (Table S13, Figures S11, S14 and S15). Nevertheless, we validated the decrease of *Dorea* in the stool of patients with N1–2 stage tumours ($p = 0.00011$) in an independent dataset (Feng et al.) (Text S2, Table S14).

3.4. Tumour CRC Microbial Subtypes

We continued our characterisation of tumour microbial heterogeneity by performing hierarchical clustering of patients based on the relative abundance of the 57 tumour genera in the tumour mucosa samples (See Methods). Once the subtypes were identified, we performed between-subtype differential abundance analyses of microbiome profiles in all three sampling environments.

Based on the tumour genera profiles, we observed three major subtypes of tumours (TMS1–TMS3), that could further be divided into two groups each (Figures 3 and 4). The bacteria were clustered into six groups B1–B6 (Figure 3, Table S12).

The B1 group and B2 group are represented by typical gut microbiome members. The B1 group consists of the five most common and most abundant genera *Fusobacterium*, *Lachnoclostridium, Bacteroides, Escherichia-Shigella* and one uncultured genus from the family *Lachnospiraceae.* All tumours contain at least three of these bacteria, most tumours (78.7%) all five. These bacteria have high co-occurrence across the sampled environments (Figure 2A, fourth panel), except for *Fusobacterium,* predominantly found in mucosa samples. The B4 group contains exclusively oral microbiome genera, and we named it the *Selenomonas* group due to its enrichment in the *Selenomonas* genera. B3 and B5 groups include mostly oral microbiome genera. These genera have significantly different incidence across the sampled environments, with 45.7–94.1% of patients missing these genera in the stool if present on tumour mucosa. Group B6 consists of 27 less common species with incidence ranging from 0% to 37% (median 11.1%).

Tumour microbial subtype 1 (TMS1) represents 26% (46) of tumours and is defined by the presence of B1–B4 microbial groups, and overall contains most of the high-grade associated genera *(Fusobacterium, Campylobacter, Leptotrichia, Peptoclostridium* and *Selenomonas,* see Table 2*).* This subtype is enriched in right-sided (60.9%), grade 3 (58.7%), pT3 or pT4 stage (95.6%) tumours and is depleted of stage 0 and stage I tumours (0% and 4.3%, respectively) (Table 1, Figure 4). In addition, TMS1 contains significantly more tumours with MSI-H (34.8%) and BRAF mutation (15.2%) compared to other tumour microbial subtypes. TMS1 differs from TMS2 and TMS3 by the presence of the *Selenomonas* group *(B4), Solobacterium* and *Howardella* species, and *Clostridium sensu stricto 1.* In contrast to other subtypes, this subtype shows a significantly decreased abundance of typical faecal commensals such as *Bifidobacterium, Ruminococcus 2, Anaerostipes* and *Coprococcus 1* on tumour mucosa (Table S15). In stool samples, we observed a higher abundance of *Prevotella* and *Clostridium sensu stricto 1* (Table S16).

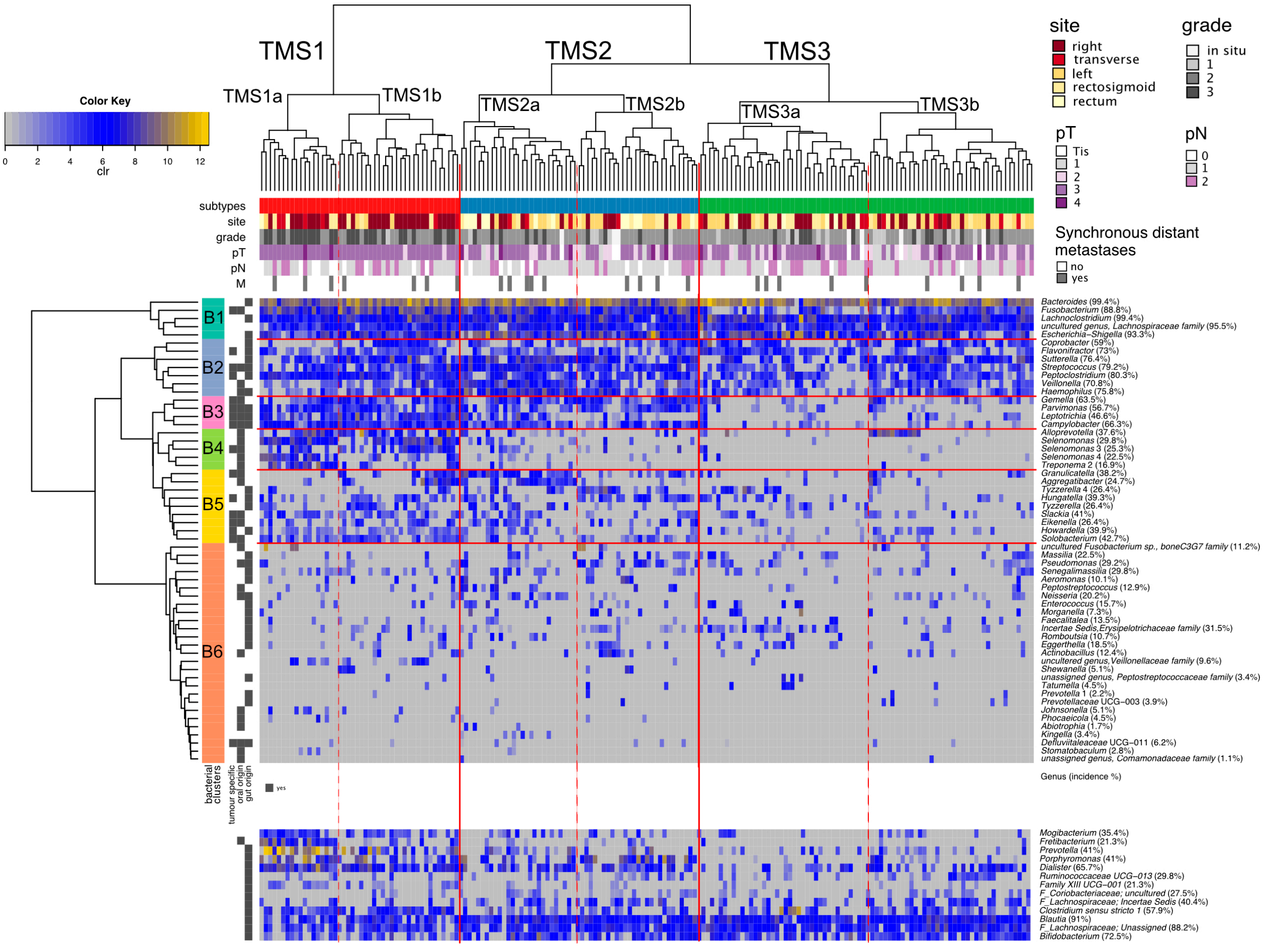


Figure 3. Tumour microbial subtypes. Hierarchical clustering of tumours (Aitchinson distance) and genera (Euclidean distance) based on the clr-transformed abundances of 57 tumour genera. Clinical variables of individual patients are shown. Proportions right to the genera name denote incidence of the genus in 178 tumour-mucosa samples. (TMS—tumour microbial subtypes, clr—centered log-ratio transformation, pT—tumour pathologic stage, pN—regional lymph nodes pathologic, M—synchronous distant metastasis, Tis—tumour in situ).

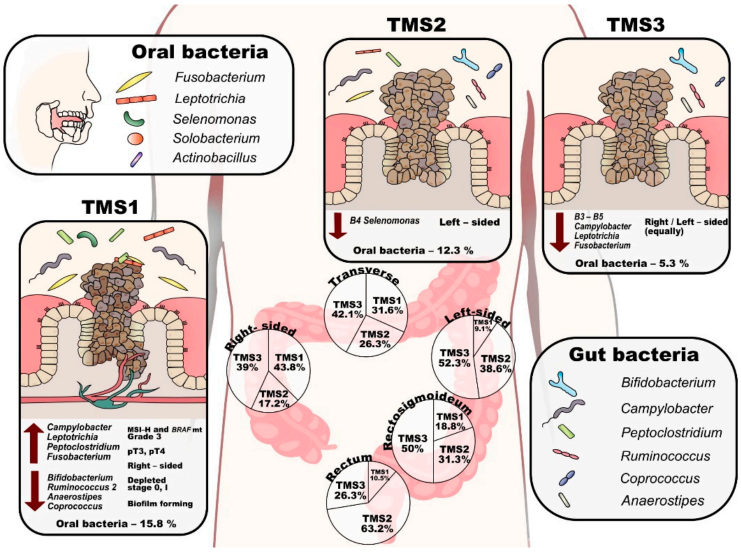


Figure 4. Scheme of the tumour microbial subtypes. (TMS—tumour microbial subtypes, pT—tumour pathologic stage, MSI-H—microsatellite instability-high).

Tumour microbial subtype 1 (TMS1) represents 26% (46) of tumours and is defined by the presence of B1–B4 microbial groups, and overall contains most of the high-grade associated genera *(Fusobacterium, Campylobacter, Leptotrichia, Peptoclostridium* and *Selenomonas,* see Table 2*).* This subtype is enriched in right-sided (60.9%), grade 3 (58.7%), pT3 or pT4 stage (95.6%) tumours and is depleted of stage 0 and stage I tumours (0% and 4.3%, respectively) (Table 1, Figure 4). In addition, TMS1 contains significantly more tumours with MSI-H (34.8%) and BRAF mutation (15.2%) compared to other tumour microbial subtypes. TMS1 differs from TMS2 and TMS3 by the presence of the *Selenomonas* group *(B4), Solobacterium* and *Howardella* species, and *Clostridium sensu stricto 1*. In contrast to other subtypes, this subtype shows a significantly decreased abundance of typical faecal commensals such as *Bifidobacterium, Ruminococcus 2, Anaerostipes* and *Coprococcus 1* on tumour mucosa (Table S15). In stool samples, we observed a higher abundance of *Prevotella* and *Clostridium sensu stricto 1* (Table S16).

Tumour microbial subtype 2 (TMS2) comprises 31% (55) of tumours and is defined mainly by the absence of B4 bacteria (the *Selenomonas* group). This subtype can be further divided into two groups by the increased incidence of *Leptotrichia, Granulicatella, Aggregatibacter* and *Neisseria* (TMS2a) or *Tyzzerella 4, Hungatella (*ex-*Clostridium hathewayi), Solobacterium, Pseudomonas* and *Porphyromonas* (TMS2b). TMS2 tumours are predominantly from the left side, rectosigmoid or rectum (70.9%) (Table 1, Figure 4). The mucosa of TMS2 tumours shows a significantly higher abundance of *Haemophilus*, *Sutterella*, *Veillonella* and *Streptococcus* and a lower abundance of *Alloprevotella* (Table S15). *Alloprevotella* was also significantly decreased in stool samples (Table S16).

Finally, the largest subtype, TMS3, represents 43% (77) of tumours and is mostly missing the B3–5 bacterial groups and most of the high-grade related species. TMS3 is characteristic by an increased proportion of grade 1 tumours (15.6%). In the TMS3 microbial subtype, right-sided and left-sided tumours are equally represented (Table 1, Figure 4). The subtype can be further divided by increased incidence of *Incertae sedis* from the *Erysipelotrichaceae* family and *Tyzzerella 4* (TMS3a) or *Clostridium sensu stricto 1, Ruminococcaceae UCG-013* and *Incertae sedis* from *Lachnospiraceae* family (TMS3b). Interestingly, subtype TMS3 contains all the tumours that lack *Fusobacterium* species (most of them in TMS3a) both in their mucosa and in the patients' stool.

Most importantly, the subtypes differed significantly in the proportion of the oral genera with TMS1 median of 15.8%, TMS2 median of 12.3% and TMS3 median of 5.3% ($p < 0.001$). We then explored the estimated metabolic potential of the microbial communities specific to the tumour subtypes (Table S17). TMS1 is characterised by functional shifts in bacterial composition, including the increase in gene content specific for nucleotide metabolism, metabolism of terpenoids and polyketides and energy metabolism ($p < 0.05$, FDR < 0.1), reduction in lipid metabolism and xenobiotics biodegradation and metabolism ($p < 0.05$, FDR < 0.1). At the lower functional level, TMS1 subtype was characterised by increased peptidoglycan biosynthesis, novobiocin biosynthesis and ansamycin biosynthesis ($p < 0.05$, FDR < 0.1). The TMS2 subtype exhibited the highest score of xenobiotics biodegradation and metabolism. TMS3 subtype showed enhanced biosynthesis of other secondary metabolites, carbohydrate metabolism and amino acid metabolism and reduced metabolism of other amino acids ($p < 0.05$, FDR < 0.1).

4. Discussion

Carcinogenesis of colorectal cancer is a complex process with a unique set of somatic molecular changes. Considerable efforts have been dedicated to understanding the heterogeneity of CRC and deriving clinically applicable molecular markers of the disease progression and patients' response to therapy. Approaching the problem from the molecular perspective in supervised analyses led to identifying several molecular markers and signatures with limited clinical use [73]. The unsupervised approach led to the definition of four consensus molecular subtypes [74], which, surprisingly, bear some prognostic value. The CRC heterogeneity puzzle, however, is far from being solved. One of the reasons is that the tumour microenvironment, especially the microbiome, seems to play a much more critical role than imagined. Many studies correlated the dysbiosis of the gut microbiome with the development of colorectal cancer in the healthy mucous-adenoma-carcinoma sequence or focused on elucidating the concrete role of selected bacterial species in the gut (colorectal) pathogenesis progression [4–22].

In contrast, our study aimed to use an unsupervised approach to characterise the heterogeneity of the CRC gut microbiome in the ongoing disease to discover unforeseen patterns. The comparison of microbial communities of stool, tumour mucosa and adjacent visually normal mucosa provided us with insights into the preferred environment of the observed species. The resulting microbial categorisation served to focus downstream analyses and to interpret our findings. We based the characterisation of the CRC tumour microbial landscape by performing subtyping of patients on bacteria with increased abundance in tumour mucosa compared to the stool to filter out potential stool contaminants. With a median of 59.6%, the tumour genera represented an essential fraction of the total microbiome found on tumour mucosa. Interestingly, the tumours of different microbial subtypes differed in the on-mucosa abundance of typical faecal genera (both pathogenic and commensal) that were not used for their definition. The analysis of microbial composition between sample types confirmed previously reported observations [8,11,34] that mucosa-associated bacteria dominate the tumour mucosal microbiome and that these species are associated also with visually normal mucosa. It is debatable to what extent the non-cancerous tissue (however distant from the tumour) from the surgically removed segment is already influenced by the bacteria initiating CRC development. Consistent with the bacterial driver-passenger model as proposed by Tjalsman et al. [27] our mucosa genera could be bacterial drivers, while tumour-specific genera could be bacterial passengers. We observed that tumours harbour a diverse community of opportunistic pathogens of oral origin (31 of 57 tumour genera) as previously reported [29,31,75].

Multiple factors make the CRC tumour niche a favourable environment for oral bacteria, in particular, for oral pathogens. Some of the bacteria can bind to specific proteins overexpressed on tumour cells [76–78]. Inflammation in the oral tissue niche selects for those species that are most adapted to the new environment, producing specific molecules such as microbial proteolytic enzymes [79] that break down the host's extracellular matrix and soluble factors to get nutrients and invade the tissue. In the digestive tract, some oral bacteria can change their oxygen requirements from facultative anaerobic to strict anaerobic and their metabolism from asaccharolytic to proteolytic [80]. Oral pathogens gaining a more favourable niche on colon tissue may shift the balance on their behalf, producing proteins playing a key role in biofilm formation [81]. Some of the oral genera detected in our study were previously never associated with CRC tumour mucosa (e.g., *Selenomonas 3, Selenomonas 4, Aggregatibacter, Johnsonella, Abiotrophia, Defluviitaleaceae UCG-011*). Most importantly, we newly associate 22 genera of both oral and gut origin with CRC overall. Some of these genera contain species that are known human pathogens causing infections of mucosal or other tissues such as: periodontal disease (*Selenomonas, Phocaeiola Aggregatibacter)* [82–84] infections in humans through animal bites (*Bergeyella* and *Actinobacillus)* [85,86]; endocarditis (*B. cardium*) or respiratory infections (*A. hominis)* [87,88]. For other genera, their potential involvement in CRC is not so obvious. The tumour-specific genera of *Defluviitaleaceae* might influence CRC through the metabolism of butyrate [89]. The association of *Tyzzerella 4* from the *Lachnospiraceae* family with CRC may be due to its increased occurrence in patients with higher cardiovascular risk (CVR) factor scores [90], which are also associated with CRC [91]. *Massilia* was detected in patients with pancreatic cancer [92].

Correlating the tumour microbiome with clinical variables of tumour progression, such as grade or stage, bears the promise of offering viable hypotheses on the role of bacteria in the progression of the disease. Currently, the associations between clinical variables and gut microbial composition in an ongoing disease are understudied, and only a few efforts addressed the topic on limited cohorts.

The sample size of our study allowed for the study of the interaction of grade and tumour location, thus providing a finer estimation of the differences in the microbiome composition. We report 59 associations of 43 genera with tumour grade and/or location for all sample types studied. We confirmed previously reported high-grade tumour associations of *Fusobacterium, Campylobacter* and *Mogibacterium,* in CRC tumour mucosa [40,68]. We newly observed potentially beneficial effects of the increased abundance of 13 *stool genera* significantly associated with left location, namely decreasing tumour grade with increasing abundance, e.g., of *Bifidobacterium, Ruminococcaceae UCG-010* and *Victivallis* in tumour mucosa; and of *Porphyromonas and Lachnospiraceae UCG-005,* in the stool. *Bifidobacterium* was previously shown to have anti-cancerogenic effects [66,69–72].

We also found location-dependent grade-predictive genera. Remarkably, while in the right colon a higher grade was associated with an increase in pathogenic genera *(Prevotella, Selenomonas)* on tumour mucosa, in the left colon a higher grade was associated with a depletion in possibly beneficial (commensal) genera *(Methanobrevibacter, Coprococcus 2, Ruminiclostridium 6, Odoribacter, Dielma, Victivallis)* on tumour mucosa or in the stool. We can only speculate whether the prolonged exposure of tumour mucosa to predominantly stool bacteria that is mechanistically related to tumours in a distant part of the colon (left-sided or with onset in rectosigmoid and rectum) can have potentially harmful or beneficial effects or whether any associations are mostly due to the well-known molecular differences in the right vs left-sided tumours [36,93,94]. The increased abundance of pathogens on the high-grade right-sided tumours might be the result of increased permeability of proximal gut mucosa [95], but the relevance of the animal models was questioned [96].

We confirmed a previously published increase of *Akkermansia* and *Porphyromonas* in the stool of patients with local metastases [72], and newly associated increased *Akkermansia* in stool in patients with stage III–IV CRC. We noted that the occurrence of synchronous local and distant metastases was mainly associated with shifts in stool microbiome, while tumour specific variables such as grade or location were associated with changes in tumour microbiome. On one side, this observation raises the possibility of microbiome-based non-invasive metastasis diagnostics in colorectal cancer or monitoring the patients at risk. On the other hand, the alteration of the stool microbial community might only reflect changes in the overall health status in the presence of metastasis and cancer progression itself similarly to non-colonic malignancies [97–99].

Pairwise analysis of the incidence of all genera across sample types helped us to assess their screening potential. On-tumour microbes with significant clinical associations and no difference in incidence across sample types are perfect candidates for stool-based screening studies or stool-based prognostic and predictive classifiers. Most of the tumour-specific genera, if present on tumour mucosa, were not identified in stool of the same patients in more than 50% of cases. Given that these genera prefer the mucosal environment over the stool, such associations are not entirely surprising. Consequently, these genera are better candidates for colonoscopy biopsy sample screening.

We then compared how the previously suggested stool-based predictive microbial markers of CRC (compared to healthy and adenomas) [29] behave with respect to the progression of an ongoing disease (associations with grade or stage) as a result of tumour microenvironment changes. Some retained their predictive potential of progression of the disease as stool predictors of the presence of local metastases (increase in *Campylobacter, Porphyromonas, Streptococcus* and decrease in *Lachnospiraceae* and *Faecalibacterium*). Some showed no significant clinical associations in stool, but their increased abundance on tumour mucosa was predictive of high pT stage (*Parvimonas*) or grade (*Fusobacterium*).

The three tumour–mucosa-based microbial subtypes we derived on patterns of similarity of the abundance of the tumour genera represent the first attempt to systematically describe microbial heterogeneity of CRC tumour environment. We were intrigued to see that compared to subtyping efforts based on gene expression [74,100], also microbial profiling identified one subtype (TMS1) enriched in *BRAF* mutant, MSI-H, right-sided tumours. An interesting observation was that the tumour microbial subtypes differ not only in the type of the tumour genera they host but also in the count of potentially pathogenic microbiome correlated with high grade and stage and the proportion of oral pathogens within the tumour genera. Of the 10 high grade or high stage-related genera, TMS1 tumours had a median of 8 (80%), TMS2 of 6 (60%) and TMS3 of 4 (40%), differing thus in what we could call "microbial pathogens burden". This subtyping could reflect differences in tumour biological properties linked with cancer progression: malignant tumours with active growth, cell and tissue atypia because of disruption of the mucus layer and dysregulation of local immunity provide more comfortable conditions to aggressive microbial consortia expansion and unconventional (oral) species homing. Moreover, with respect to the bacteria-supported model of carcinogenesis, proved in animal models [10,101], the pathogenic bacteria growth leads to additional dedifferentiation of tumour cells forming the pathogenetic loop. The differences in the proportion of oral pathogens and metabolic potential lead us to the hypothesis that the TMS1 subtype is enriched in tumours with microbial biofilms. This subtype is enriched in right-sided tumours and compared with the other two subtypes, is enriched in the presence of oral bacteria. Recently, biofilms have been associated with right-sided CRC [31]. Drewes et al. [102] identified several biofilm-associated shifts, including the functional alteration in peptidoglycan biosynthesis, novobiocin biosynthesis and ansamycin biosynthesis, which were significantly increased in the TMS1 subtype. Studies show that the commensal and the pathogenic periodontal bacteria *(Fusobacterium, Porphypomonas)* produce proteins such as gingipains [81] and RadD [103], which can play a key role in biofilm formation. Koliarakis et al. [75] proposed a new outlook on CRC pathogenesis driven by gut mucosa biofilm created by periodontopathic bacteria translocated into the colorectum. Tomkovich et al. [104] successfully demonstrated that polymicrobial biofilms are carcinogenic. Transcriptomic studies of periodontal tissues show that many organisms can fulfil gaps in metabolism, therefore the pathogenic community is more important as one unit than the virulence of one species [105].

It remains to be investigated whether the microbial subtypes could improve the prediction of patients' survival and prognosis. We can speculate that the high microbial pathogen burden could worsen not only the tumour progression but also potentially the patient's condition after the surgical resection and during and after the chemotherapy treatment. Given the fact that tumour-related genera reside also on visually normal mucosa, they could initiate CRC tissue dysplastic changes and malignisation. There is limited evidence of linkage between mucosal microbiota and metachronous adenomas growing demonstrated by Liu et al. [106]. On the other side, it is shown that the microbiome could interact and metabolise chemotherapeutic medicine, which leads to modulation of its activity and toxicity [107]. In the light of the above, modification of gut microbiome after colorectal cancer surgical removal might be considered as an additional step of treatment to prevent tumour recurrence and modulate chemotherapy effectiveness and toxicity.

The probable presence of biofilm in the TMS1 subtype might make this subtype of interest to potential prevention and treatment strategies. Importantly, although TMS1 is enriched in proximal tumours, it occurs in 9.1–18.8% of left, rectosigmoid and rectal tumours. Remarkably, biofilm communities from the colon biopsies of healthy individuals were as potent in inducing colon inflammation as the biofilm communities from CRC hosts [104]. The inhibition or removal of such biofilms from patients with CRC could represent a promising strategy for secondary CRC prevention and treatment but remains an uneasy task due to inefficiency of traditional antimicrobial strategies such as antibiotic treatment [108]. A recent study associated periodontitis with increased risk of high-grade proximal colorectal cancer [109]. Based on this, our results suggest an intriguing hypothesis: whether improving oral hygiene would impact the incidence of TMS1 tumours, or, more importantly, would lower the recurrence rate or development of secondary tumours in the TMS1 patients.

Another clinically relevant observation is the association of left-sided high-grade tumours with the depletion in protective species rather than increase in bacterial pathogens. This suggests that antibiotic treatment of patients with distal tumours may have a detrimental effect on their prognosis. Coadministration of probiotics in this case could be highly beneficial.

We believe that for certain patient populations, the inclusion of tumour mucosa sampling during colonoscopy for analysis of microbial composition could help to efficiently steer pre- and post-operative treatment decisions.

5. Conclusions

In our study, by analysis of 483 samples from n = 178 patients, we extended the current characterisation of the colorectal cancer microbiome in several directions. Thanks to the large sample size, we identified bacterial genera that were not previously associated with CRC tumour mucosa, clinical variables or with colorectal carcinoma at all. These genera should be studied in more detail to describe their mechanism of interaction with the disease.

By focusing on microbial community analysis, in contrast to classical microbiome-centred approaches, we were able to identify co-occurring species and three major tumour-microbial subtypes that correlate with clinical variables, such as grade, location and TNM staging. The subtypes also differ in what we describe as microbial pathogens burden—the number of pathogenic species correlated with increased grade and stage present on tumour mucosa, although the concept can be defined with respect to all three environments (tumour mucosa, visually normal mucosa and stool).

An important limitation of our study is the lack of proper validation of all the results since adequate data is unavailable, hence these results must be taken cautiously. Additionally, it is well known that the gut microbial composition changes with dietary patterns and lifestyle, which could be region-based [98]. More studies of similar sample size or larger from different geographical locations, are needed to derive robust and generalisable patterns. We make the full data available, including clinical variables, as a first step towards building a data corpus that could support such investigations. The nature of the samples (mucosa) prohibited us from using more advanced whole-metagenomic sequencing due to severe human DNA contamination issues [110–112]. The technology chosen was high throughput, fitting the purpose of microbial community-based analysis. We did perform the sequence matching for the identified ASVs against the SILVA database, however, being aware of the limitations, we provide these results solely as supplementary information without discussing them here in detail.

Having sampled the microbiome at three different complementary sites allowed the study of several environments leading to the definition of novel microbial categories with multiple implications. Our study shows that the associations with clinical variables found for the tumour mucosal or adjacent visually normal mucosa microbiome are rarely preserved in the microbial composition of stool and vice versa. While tumour histological grade, stage and location are reflected in the corresponding mucosal microbiome, the presence of lymph nodes or distant metastases influences mainly the stool microbiome. It seems that the mucosa and stool microbiome are complementary with respect to the modulation of their effects on disease progression. Tumour-mucosa biopsies from colonoscopy might need to be coupled with stool sampling for efficient screening or diagnostic purposes.

Understanding the role of tumour-subtype specific microbial communities could lead to tailored strategies of CRC patient gut microbiome management through lifestyle and diet recommendations including probiotic and antimicrobial interventions.

Our study is a step forward in understanding the role of the microbiome and its interactions with other factors involved in oncogenesis and tumour progression. Rather than providing definite answers it opens new avenues for exploring new treatments and biomarkers.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/cancers13194799/s1, Figure S1: Number of reads and diversity comparison of three sample types on 127 triplets, Figure S2: Genera which were detected only in some of the three sample types, Figure S3: Differences in microbiome composition across the sample types, Figure S4: Co-occurrence analysis of 57 tumour genera, Figure S5: Microbial β-diversity analysis by NMDS on 178 tumour tissue swab, Figure S6: Microbial β-diversity analysis by NMDS performed on 127 triplets, Figure S7: Side-dependent associations between tumour histological grade and microbiota composition, Figure S8: Boxplots of distribution of clr transformed abundance of genera associated with tumour grade and/or location in 178 tumour mucosa samples in models with or without interaction at *p*-value < 0.05, Figure S9: Boxplots of distribution of clr transformed abundance of genera associated with tumour grade and/or location in 178 visually normal mucosa samples in models with or without interaction at *p*-value < 0.05, Figure S10: Boxplots of distribution of clr transformed abundance of genera associated with tumour grade and/or location in 127 stool samples in models with or without interaction at *p*-value < 0.05, Figure S11: Associations between tumour stage, including TNM staging separately, gender and microbiota composition in all sample types and sample sizes, Figure S12: Boxplots of distribution of clr transformed abundance of genera associated with tumour stage in 178 tumour mucosa samples, 178 adjacent visually normal mucosa samples and 127 stool samples at *p*-value < 0.05, Figure S13: Boxplots of distribution of clr transformed abundance of genera associated with tumour pathologic stage in 178 tumour mucosa samples, 178 adjacent visually normal mucosa samples and 127 stool samples at *p*-value < 0.05, Figure S14: Boxplots of distribution of clr transformed abundance of genera associated with the presence of lymph-node metastases in 178 tumour mucosa samples, 178 adjacent visually normal mucosa samples and 127 stool samples at *p*-value < 0.05, Figure S15: Boxplots of distribution of clr transformed abundance of genera associated with the presence of distance metastases in 178 tumour mucosa samples, 178 adjacent visually normal mucosa samples and 127 stool samples at *p*-value < 0.05, Figure S16: Results of pairwise coincidence analysis of genera across sample types within the same patient, Table S1: Results of validation of associations of microbiome with tumour location on the Dejea et al. dataset, Table S2: Results of validation of associations of microbiome with tumour location on the Dejea et al., Zeller et al. and Feng et al. datasets, Table S3: List of primers and the length of PCR products, Table S4: Results of the filtering steps based on ASV abundance and type of taxonomic assignment, Table S5: Table of identified ASV's in the three sample types (127 triplets) with taxonomy assigned by BLAST and QIIME, Table S6: Number of taxa identified at respective taxonomic levels after the filtering steps in 127 triplets (381 samples) and overall (483), in different sample types and their combinations, Table S7: Results of Cochran's Q test and McNemar's test including the report of the results as the co-occurence of specific event pairs: genera present in one sample type but not in the second sample type and vice versa, Table S8: Total counts of genera found significantly differentially abundant across the three sample types (127 triplets), divided into categories according to their enrichment in different sample types (TtoS—tumour to stool, VNtoS—visually-normal to stool, TtoVN—tumour to visually-normal.), Table S9: Results of the Friedman test across the three sample types (127 triplets), Table S10: Results of microbial co-occurrence analysis, Table S11. Summary of the 57 tumour genera. Overview of the current knowledge about the genera that we categorise as associated with tumour mucosa, Table S12: Results of adonis testing of the associations between β-diversity and clinical variables, Table S13: Results of rank regression associating microbiome with the clinical variables, Table S14: Results of rank regression on publicly available validation datasets, Table S15: Results of rank regression associating tumour microbiome abundance with tumour microbiome subtypes, Table S16: Results of rank regression associating stool microbiome abundance with tumour microbiome subtypes, Table S17: Results of rank regression of differences between tumour microbiome subtypes in metabolic potential of the microbial communities, Text S1: Extended results of differences in microbiome diversity and incidence across the sample type, Text S2: Validation of results on publicly available data.

Author Contributions: E.B., R.Š., R.N., B.B. and L.Z.-D. designed the study; B.Z. and S.S. performed bioinformatic analysis; B.Z., V.A.P., V.P. and E.B. performed statistical data analysis; M.H., L.M., N.K. and V.B. performed the sample processing, DNA isolation and sequencing; B.Z., M.H., V.A.P., L.M., E.B., P.V., L.Z.-D., B.B. and R.H. interpreted the results; B.Z. drafted the first version of the manuscript. M.H. designed Figures 1 and 4. All authors participated in writing or editing of the manuscript and approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Authors thanks to Research Infrastructure RECETOX RI (No LM2018121) financed by the Ministry of Education, Youth and Sports, and Operational Programme Research, Development and Innovation—project CETOCOEN EXCELLENCE (No CZ.02.1.01/0.0/0.0/17_043/0009632) for supportive background. This work received funding from the Ministry of Health of the Czech Republic (project AZV 16-31966A) and the European Union's Horizon 2020 research and innovation programme under grant agreement No 825410. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development, and Innovations Infrastructures. The work was supported by the project BBMRI-CZ no. LM2018125.

Institutional Review Board Statement: The study was approved by the ethical committee of Masaryk Memorial Cancer Institute. Ethical code: Č.J. 2015/1721/MOU, JID: MOU 73 866.

Informed Consent Statement: All patients gave written informed consent in accordance with the Declaration of Helsinki prior to participating in the study.

Data Availability Statement: Sequencing data were uploaded to the European Nucleotide Archive under accession number PRJEB35990.

Acknowledgments: The authors acknowledge the help of the volunteers enrolled in this study. We thank the ONCOBIOME researchers for their insight and suggestions.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Dyba, T.; Randi, G.; Bettio, M.; Gavin, A.; Visser, O.; Bray, F. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer* **2018**, *103*, 356–387. [CrossRef]
2. Punt, C.J.A.; Koopman, M.; Vermeulen, L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 235–246. [CrossRef]
3. Van Der Jeught, K.; Xu, H.-C.; Li, Y.-J.; Lu, X.-B.; Ji, G. Drug resistance and new therapies in colorectal cancer. *World J. Gastroenterol.* **2018**, *24*, 3834–3848. [CrossRef]
4. Ahn, J.; Sinha, R.; Pei, Z.; Dominianni, C.; Wu, J.; Shi, J.; Goedert, J.J.; Hayes, R.; Yang, L. Human Gut Microbiome and Risk for Colorectal Cancer. *J. Natl. Cancer Inst.* **2013**, *105*, 1907–1911. [CrossRef] [PubMed]
5. Arthur, J.; Perez-Chanona, E.; Mühlbauer, M.; Tomkovich, S.; Uronis, J.M.; Fan, T.-J.; Campbell, B.J.; Abujamel, T.; Dogan, B.; Rogers, A.B.; et al. Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota. *Science* **2012**, *338*, 120–123. [CrossRef] [PubMed]
6. Balamurugan, R.; Rajendiran, E.; George, S.; Samuel, G.V.; Ramakrishna, B. Real-time polymerase chain reaction quantification of specific butyrate-producing bacteria, DesulfovibrioandEnterococcus faecalisin the feces of patients with colorectal cancer. *J. Gastroenterol. Hepatol.* **2008**, *23*, 1298–1303. [CrossRef] [PubMed]
7. Chen, J.; Young, S.M.; Allen, C.; Seeber, A.; Péli-Gulli, M.-P.; Panchaud, N.; Waller, A.; Ursu, O.; Yao, T.; Golden, J.E.; et al. Identification of a Small Molecule Yeast TORC1 Inhibitor with a Multiplex Screen Based on Flow Cytometry. *ACS Chem. Biol.* **2012**, *7*, 715–722. [CrossRef] [PubMed]
8. Chen, W.; Liu, F.; Ling, Z.; Tong, X.; Xiang, C. Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer. *PLoS ONE* **2012**, *7*, e39743. [CrossRef] [PubMed]
9. Cipe, G.; Idiz, U.O.; Firat, D.; Bektasoglu, H. Relationship between intestinal microbiota and colorectal cancer. *World J. Gastrointest. Oncol.* **2015**, *7*, 233–240. [CrossRef]
10. Kostic, A.; Chun, E.; Robertson, L.; Glickman, J.N.; Gallini, C.A.; Michaud, M.; Clancy, T.E.; Chung, D.C.; Lochhead, P.; Hold, G.; et al. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host Microbe* **2013**, *14*, 207–215. [CrossRef]
11. Lu, Y.; Chen, J.; Zheng, J.; Hu, G.; Wang, J.; Huang, C.; Lou, L.; Wang, X.; Zeng, Y. Mucosal adherent bacterial dysbiosis in patients with colorectal adenomas. *Sci. Rep.* **2016**, *6*, 26337. [CrossRef]
12. Marchesi, J.R.; Dutilh, B.E.; Hall, N.; Peters, W.H.M.; Roelofs, R.; Boleij, A.; Tjalsma, H. Towards the Human Colorectal Cancer Microbiome. *PLoS ONE* **2011**, *6*, e20447. [CrossRef]
13. Nakatsu, G.; Li, X.; Zhou, H.; Sheng, J.; Wong, S.H.; Wu, W.K.K.; Ng, S.C.; Tsoi, H.; Dong, Y.; Zhang, N.; et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* **2015**, *6*, 8727. [CrossRef] [PubMed]
14. Rubinstein, M.R.; Wang, X.; Liu, W.; Hao, Y.; Cai, G.; Han, Y.W. Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. *Cell Host Microbe* **2013**, *14*, 195–206. [CrossRef]
15. Sobhani, I.; Tap, J.; Roudot-Thoraval, F.; Roperch, J.P.; Letulle, S.; Langella, P.; Corthier, G.; Van Nhieu, J.T.; Furet, J.-P. Microbial Dysbiosis in Colorectal Cancer (CRC) Patients. *PLoS ONE* **2011**, *6*, e16393. [CrossRef]
16. Viljoen, K.S.; Dakshinamurthy, A.; Goldberg, P.; Blackburn, J.M. Quantitative Profiling of Colorectal Cancer-Associated Bacteria Reveals Associations between Fusobacterium spp., Enterotoxigenic Bacteroides fragilis (ETBF) and Clinicopathological Features of Colorectal Cancer. *PLoS ONE* **2015**, *10*, e0119462. [CrossRef]

17. Wang, T.; Cai, G.; Qiu, Y.; Fei, N.; Zhang, M.; Pang, X.; Jia, W.; Cai, S.; Zhao, L. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* **2011**, *6*, 320–329. [CrossRef]
18. Wu, N.; Yang, X.; Zhang, R.; Li, J.; Xiao, X.; Hu, Y.; Chen, Y.; Yang, F.; Lu, N.; Wang, Z.; et al. Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients. *Microb. Ecol.* **2013**, *66*, 462–470. [CrossRef] [PubMed]
19. Yang, J.; McDowell, A.; Kim, E.K.; Seo, H.; Lee, W.H.; Moon, C.-M.; Kym, S.-M.; Lee, D.H.; Park, Y.S.; Jee, Y.-K.; et al. Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis. *Exp. Mol. Med.* **2019**, *51*, 1–15. [CrossRef] [PubMed]
20. Yu, J.; Feng, Q.; Wong, S.H.; Zhang, D.; Liang, Q.Y.; Qin, Y.; Tang, L.; Zhao, H.; Stenvang, J.; Li, Y.; et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **2017**, *66*, 70–78. [CrossRef] [PubMed]
21. Zackular, J.P.; Baxter, N.; Iverson, K.D.; Sadler, W.D.; Petrosino, J.F.; Chen, G.Y.; Schloss, P.D. The Gut Microbiome Modulates Colon Tumorigenesis. *mBio* **2013**, *4*, e00692-13. [CrossRef]
22. Zackular, J.; Rogers, M.; Ruffin, M.; Schloss, P.D. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. *Cancer Prev. Res.* **2014**, *7*, 1112–1121. [CrossRef]
23. Vaupel, P.; Harrison, L. Tumor Hypoxia: Causative Factors, Compensatory Mechanisms, and Cellular Response. *Oncologist* **2004**, *9* (Suppl. S5), 4–9. [CrossRef] [PubMed]
24. Louis, P.; Hold, G.; Flint, H.J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **2014**, *12*, 661–672. [CrossRef] [PubMed]
25. Tlaskalova-Hogenova, H.; Vannucci, L.; Klimesova, K.; Stepankova, R.; Krizan, J.; Kverka, M. Microbiome and Colorectal Carcinoma. *Cancer J.* **2014**, *20*, 217–224. [CrossRef] [PubMed]
26. Xiao, Y.; Freeman, G.J. The Microsatellite Instable Subset of Colorectal Cancer Is a Particularly Good Candidate for Checkpoint Blockade Immunotherapy. *Cancer Discov.* **2015**, *5*, 16–18. [CrossRef]
27. Tjalsma, H.; Boleij, A.; Marchesi, J.R.; Dutilh, B.E. A bacterial driver–passenger model for colorectal cancer: Beyond the usual suspects. *Nat. Rev. Genet.* **2012**, *10*, 575–582. [CrossRef]
28. Pennisi, E. Cancer Therapies Use a Little Help from Microbial Friends. *Science* **2013**, *342*, 921. [CrossRef]
29. Thomas, A.M.; Manghi, P.; Asnicar, F.; Pasolli, E.; Armanini, F.; Zolfo, M.; Beghini, F.; Manara, S.; Karcher, N.; Pozzi, C.; et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **2019**, *25*, 667–678. [CrossRef]
30. Feng, Q.; Liang, S.; Jia, H.; Stadlmayr, A.; Tang, L.; Lan, Z.; Zhang, D.; Xia, H.; Xu, X.; Jie, Z.; et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **2015**, *6*, 6528. [CrossRef]
31. Dejea, C.M.; Wick, E.C.; Hechenbleikner, E.M.; White, J.R.; Welch, J.L.M.; Rossetti, B.J.; Peterson, S.N.; Snesrud, E.C.; Borisy, G.G.; Lazarev, M.; et al. Microbiota organization is a distinct feature of proximal colorectal cancers. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 18321–18326. [CrossRef] [PubMed]
32. Zeller, G.; Tap, J.; Voigt, A.Y.; Sunagawa, S.; Kultima, J.R.; Costea, P.I.; Amiot, A.; Böhm, J.; Brunetti, F.; Habermann, N.; et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **2014**, *10*, 766. [CrossRef]
33. Yang, Y.; Cai, Q.; Shu, X.; Steinwandel, M.D.; Blot, W.J.; Zheng, W.; Long, J. Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations. *Int. J. Cancer* **2019**, *144*, 2381–2389. [CrossRef]
34. Liu, C.; Zhang, Y.; Shang, Y.; Wu, B.; Yang, E.; Luo, Y.-Y.; Li, X. Intestinal bacteria detected in cancer and adjacent tissue from patients with colorectal cancer. *Oncol. Lett.* **2018**, *17*, 1115–1127. [CrossRef] [PubMed]
35. Pu, L.Z.C.T.; Yamamoto, K.; Honda, T.; Nakamura, M.; Yamamura, T.; Hattori, S.; Burt, A.D.; Singh, R.; Hirooka, Y.; Fujishiro, M. Microbiota profile is different for early and invasive colorectal cancer and is consistent throughout the colon. *J. Gastroenterol. Hepatol.* **2020**, *35*, 433–437. [CrossRef]
36. Flemer, B.; Lynch, D.B.; Brown, J.M.R.; Jeffery, I.; Ryan, F.; Claesson, M.; O'Riordain, M.; Shanahan, F.; O'Toole, P.W. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **2017**, *66*, 633–643. [CrossRef]
37. Egao, Z.; Eguo, B.; Egao, R.; Ezhu, Q.; Eqin, H. Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* **2015**, *6*, 20. [CrossRef]
38. Li, E.; Hamm, C.M.; Gulati, A.S.; Sartor, R.B.; Chen, H.; Wu, X.; Zhang, T.; Rohlf, F.J.; Zhu, W.; Gu, C.; et al. Inflammatory Bowel Diseases Phenotype, C. difficile and NOD2 Genotype Are Associated with Shifts in Human Ileum Associated Microbial Composition. *PLoS ONE* **2012**, *7*, e26284. [CrossRef]
39. Han, S.; Wu, W.; Da, M.; Xu, J.; Zhuang, J.; Zhang, L.; Zhang, X.; Yang, X. Adequate Lymph Node Assessments and Investigation of Gut Microorganisms and Microbial Metabolites in Colorectal Cancer. *OTT* **2020**, *13*, 1893–1906. [CrossRef]
40. Wu, Y.; Shi, L.; Li, Q.; Wu, J.; Peng, W.; Li, H.; Chen, K.; Ren, Y.; Fu, X. Microbiota Diversity in Human Colorectal Cancer Tissues Is Associated with Clinicopathological Features. *Nutr. Cancer* **2019**, *71*, 214–222. [CrossRef]
41. Apprill, A.; McNally, S.; Parsons, R.; Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **2015**, *75*, 129–137. [CrossRef]
42. Caporaso, J.G.; Lauber, C.L.; Walters, W.A.; Berg-Lyons, D.; Lozupone, C.A.; Turnbaugh, P.J.; Fierer, N.; Knight, R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4516–4522. [CrossRef]
43. Callahan, B.J.; McMurdie, P.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]

44. Aronesty, E. Comparison of Sequencing Utility Programs. *TOBIOIJ* **2013**, *7*, 1–8. [CrossRef]
45. Pruesse, E.; Quast, C.; Knittel, K.; Fuchs, B.M.; Ludwig, W.; Peplies, J.; Glöckner, F.O. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **2007**, *35*, 7188–7196. [CrossRef]
46. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef]
47. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.; Costello, E.K.; Fierer, N.; Peña, A.G.; Goodrich, J.K.; Gordon, J.I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335–336. [CrossRef] [PubMed]
48. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
49. Lozupone, C.; Knight, R. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* **2005**, *71*, 8228–8235. [CrossRef] [PubMed]
50. Douglas, G.M.; Maffei, V.J.; Zaneveld, J.R.; Yurgel, S.N.; Brown, J.R.; Taylor, C.M.; Huttenhower, C.; Langille, M.G.I. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **2020**, *38*, 685–688. [CrossRef] [PubMed]
51. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B Methodol.* **1982**, *44*, 139–160. [CrossRef]
52. Gloor, G.B.; Wu, J.R.; Pawlowsky-Glahn, V.; Egozcue, J.J. It's all relative: Analyzing microbiome data as compositions. *Ann. Epidemiol.* **2016**, *26*, 322–329. [CrossRef]
53. Martín-Fernández, J.-A.; Hron, K.; Templ, M.; Filzmoser, P.; Palarea-Albaladejo, J. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model. Int. J.* **2015**, *15*, 134–158. [CrossRef]
54. Tsilimigras, M.C.; Fodor, A.A. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Ann. Epidemiol.* **2016**, *26*, 330–335. [CrossRef] [PubMed]
55. Oaksen, J.; Blanchet, F.G.; Friendly, M.; Kindt, R.; Legendre, P.; McGlinn, D.; Minchin, P.R.; O'Hara, R.B.; Simpson, G.L.; Solymos, P.; et al. Vegan: Community Ecology Package. R Package. 2019. Available online: http://cran.rproject.org/package=vegan (accessed on 4 July 2020).
56. Comas-Cufí, M. R Package. coda.base: A Basic Set of Functions for Compositional Data Analysis. 2020. Available online: https://rdrr.io/cran/coda.base/ (accessed on 4 July 2020).
57. Kloke, J.D.; McKean, J.W. Rfit: Rank-Based Estimation for Linear Models. *R J.* **2012**, *4*, 57–64. [CrossRef]
58. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
59. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. *circlize* implements and enhances circular visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [CrossRef] [PubMed]
60. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
61. Warnes, G.R.; Bolker, B.; Bonebakker, L.; Gentleman, R.; Liaw, W.H.A.; Lumley, T.; Maechler, M.; Magnusson, A.; Moeller, S.; Schwartz, M.; et al. Gplots: Various R Programming Tools for Plotting Data; R Package. 2020. Available online: https://rdrr.io/cran/gplots/ (accessed on 4 July 2020).
62. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*, 2nd ed.; Springer: Cham, Switzerland, 2016; ISBN 978-3-319-24277-4.
63. Wray, C.M.; Ziogas, A.; Hinojosa, M.W.; Le, H.; Stamos, M.J.; Zell, J.A. Tumor Subsite Location Within the Colon Is Prognostic for Survival After Colon Cancer Diagnosis. *Dis. Colon Rectum* **2009**, *52*, 1359–1366. [CrossRef] [PubMed]
64. Pasolli, E.; Schiffer, L.; Manghi, P.; Renson, A.; Obenchain, V.; Truong, D.T.; Beghini, F.; Malik, F.; Ramos, M.; Dowd, J.; et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **2017**, *14*, 1023–1024. [CrossRef]
65. De Almeida, C.V.; Lulli, M.; Di Pilato, V.; Schiavone, N.; Russo, E.; Nannini, G.; Baldi, S.; Borrelli, R.; Bartolucci, G.; Menicatti, M.; et al. Differential Responses of Colorectal Cancer Cell Lines to Enterococcus faecalis' Strains Isolated from Healthy Donors and Colorectal Cancer Patients. *JCM* **2019**, *8*, 388. [CrossRef]
66. Gupta, A.; Dhakan, D.B.; Maji, A.; Saxena, R.; Visnu Prasoodanan, P.K.; Mahajan, S.; Pulikkan, J.; Kurian, J.; Gomez, A.M.; Scaria, J.; et al. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems* **2019**, *4*. [CrossRef]
67. Ai, D.; Pan, H.; Li, X.; Gao, Y.; Liu, G.; Xia, L.C. Identifying Gut Microbiota Associated with Colorectal Cancer Using a Zero-Inflated Lognormal Model. *Front. Microbiol.* **2019**, *10*, 826. [CrossRef]
68. Ito, M.; Kanno, S.; Nosho, K.; Sukawa, Y.; Mitsuhashi, K.; Kurihara, H.; Igarashi, H.; Takahashi, T.; Tachibana, M.; Takahashi, H.; et al. Association ofFusobacterium nucleatumwith clinical and molecular features in colorectal serrated pathway. *Int. J. Cancer* **2015**, *137*, 1258–1268. [CrossRef]
69. Bahmani, S.; Azarpira, N.; Moazamian, E. Anti-colon cancer activity of Bifidobacterium metabolites on colon cancer cell line SW742. *Turk. J. Gastroenterol.* **2019**, *30*, 835–842. [CrossRef]
70. Mangifesta, M.; Mancabelli, L.; Milani, C.; Gaiani, F.; De'Angelis, N.; De'Angelis, G.L.; Van Sinderen, D.; Ventura, M.; Turroni, F. Mucosal microbiota of intestinal polyps reveals putative biomarkers of colorectal cancer. *Sci. Rep.* **2018**, *8*, 1–9. [CrossRef] [PubMed]
71. Parisa, A.; Roya, G.; Mahdi, R.; Shabnam, R.; Maryam, E.; Malihe, T. Anti-cancer effects of Bifidobacterium species in colon cancer cells and a mouse model of carcinogenesis. *PLoS ONE* **2020**, *15*, e0232930. [CrossRef] [PubMed]

72. Sivan, A.; Corrales, L.; Hubert, N.; Williams, J.B.; Aquino-Michaels, K.; Earley, Z.M.; Benyamin, F.W.; Lei, Y.M.; Jabri, B.; Alegre, M.-L.; et al. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **2015**, *350*, 1084–1089. [CrossRef] [PubMed]
73. Koncina, E.; Haan, S.; Rauh, S.; Letellier, E. Prognostic and Predictive Molecular Biomarkers for Colorectal Cancer: Updates and Challenges. *Cancers* **2020**, *12*, 319. [CrossRef] [PubMed]
74. Guinney, J.; Dienstmann, R.; Wang, X.; De Reyniès, A.; Schlicker, A.; Soneson, C.; Marisa, L.; Roepman, P.; Nyamundanda, G.; Angelino, P.; et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **2015**, *21*, 1350–1356. [CrossRef] [PubMed]
75. Koliarakis, I.; Messaritakis, I.; Nikolouzakis, T.K.; Hamilos, G.; Souglakos, J.; Tsiaoussis, J. Oral Bacteria and Intestinal Dysbiosis in Colorectal Cancer. *Int. J. Mol. Sci.* **2019**, *20*, 4146. [CrossRef]
76. Long, X.; Wong, C.C.; Tong, L.; Chu, E.S.H.; Szeto, C.H.; Go, M.Y.Y.; Coker, O.O.; Chan, A.W.H.; Chan, F.K.; Sung, J.J.Y.; et al. Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.* **2019**, *4*, 2319–2330. [CrossRef]
77. Abed, J.; Emgård, J.E.; Zamir, G.; Faroja, M.; Almogy, G.; Grenov, A.; Sol, A.; Naor, R.; Pikarsky, E.; Atlan, K.A.; et al. Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe* **2016**, *20*, 215–225. [CrossRef]
78. Zou, X.; Feng, B.; Dong, T.; Yan, G.; Tan, B.; Shen, H.; Huang, A.; Zhang, X.; Zhang, M.; Yang, P.; et al. Up-regulation of type I collagen during tumorigenesis of colorectal cancer revealed by quantitative proteomic analysis. *J. Proteom.* **2013**, *94*, 473–485. [CrossRef] [PubMed]
79. Takahashi, N. Microbial ecosystem in the oral cavity: Metabolic diversity in an ecological niche and its relationship with oral diseases. *Int. Congr. Ser.* **2005**, *1284*, 103–112. [CrossRef]
80. Eley, B.M.; Cox, S.W. Proteolytic and hydrolytic enzymes from putative periodontal pathogens: Characterization, molecular genetics, effects on host defenses and tissues and detection in gingival crevice fluid. *Periodontol. 2000* **2003**, *31*, 105–124. [CrossRef] [PubMed]
81. Potempa, J.; Sroka, A.; Imamura, T.; Travis, J. Gingipains, the major cysteine proteinases and virulence factors of Porphyromonas gingivalis: Structure, function and assembly of multidomain protein complexes. *Curr. Protein Pept. Sci.* **2003**, *4*, 397–407. [CrossRef] [PubMed]
82. Gonçalves, L.F.H.; Fermiano, D.; Feres, M.; Figueiredo, L.C.; Teles, F.R.P.; Mayer, M.P.A.; Faveri, M. Levels ofSelenomonasspecies in generalized aggressive periodontitis. *J. Periodontal Res.* **2012**, *47*, 711–718. [CrossRef]
83. Scher, J.U.; Ubeda, C.; Equinda, M.; Khanin, R.; Buischi, Y.; Viale, A.; Lipuma, L.; Attur, M.; Pillinger, M.; Weissmann, G.; et al. Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum.* **2012**, *64*, 3083–3094. [CrossRef] [PubMed]
84. Herbert, B.A.; Novince, C.M.; Kirkwood, K.L. Aggregatibacter actinomycetemcomitans, a potent immunoregulator of the periodontal host defense system and alveolar bone homeostasis. *Mol. Oral Microbiol.* **2016**, *31*, 207–227. [CrossRef]
85. Lin, W.-R.; Chen, Y.-S.; Liu, Y.-C. Cellulitis and Bacteremia Caused by Bergeyella zoohelcum. *J. Formos. Med. Assoc.* **2007**, *106*, 573–576. [CrossRef]
86. Peel, M.M.; Hornidge, K.A.; Luppino, M.; Stacpoole, A.M.; Weaver, R.E. Actinobacillus spp. and related bacteria in infected wounds of humans bitten by horses and sheep. *J. Clin. Microbiol.* **1991**, *29*, 2535–2538. [CrossRef]
87. Sohn, K.M.; Huh, K.; Baek, J.-Y.; Kim, Y.-S.; Kang, C.-I.; Peck, K.R.; Lee, N.Y.; Song, J.-H.; Ko, K.S.; Chung, D.R. A new causative bacteria of infective endocarditis, Bergeyella cardium sp. nov. *Diagn. Microbiol. Infect. Dis.* **2015**, *81*, 213–216. [CrossRef]
88. Friis-Møller, A.; Christensen, J.J.; Fussing, V.; Hesselbjerg, A.; Christiansen, J.; Bruun, B. Clinical Significance and Taxonomy of Actinobacillus hominis. *J. Clin. Microbiol.* **2001**, *39*, 930–935. [CrossRef]
89. Zha, Z.; Lv, Y.; Tang, H.; Li, T.; Miao, Y.; Cheng, J.; Wang, G.; Tan, Y.; Zhu, Y.; Xing, X.; et al. An orally administered butyrate-releasing xylan derivative reduces inflammation in dextran sulphate sodium-induced murine colitis. *Int. J. Biol. Macromol.* **2020**, *156*, 1217–1233. [CrossRef]
90. Kelly, T.N.; Bazzano, L.A.; Ajami, N.J.; He, H.; Zhao, J.; Petrosino, J.F.; Correa, A.; He, J. Gut Microbiome Associates with Lifetime Cardiovascular Disease Risk Profile Among Bogalusa Heart Study Participants. *Circ. Res.* **2016**, *119*, 956–964. [CrossRef] [PubMed]
91. Niederseer, D.; Bracher, I.; Stadlmayr, A.; Huber-Schönauer, U.; Plöderl, M.; Obeid, S.; Schmied, C.; Hammerl, S.; Stickel, F.; Lederer, D.; et al. Association between Cardiovascular Risk and Diabetes with Colorectal Neoplasia: A Site-Specific Analysis. *J. Clin. Med.* **2018**, *7*, 484. [CrossRef] [PubMed]
92. Mei, Q.-X.; Huang, C.-L.; Luo, S.-Z.; Zhang, X.-M.; Zeng, Y.; Lu, Y.-Y. Characterization of the duodenal bacterial microbiota in patients with pancreatic head cancer vs. healthy controls. *Pancreatology* **2018**, *18*, 438–445. [CrossRef] [PubMed]
93. Gao, R.; Kong, C.; Huang, L.; Li, H.; Qu, X.; Liu, Z.; Lan, P.; Wang, J.; Qin, H. Mucosa-associated microbiota signature in colorectal cancer. *Eur. J. Clin. Microbiol. Infect. Dis.* **2017**, *36*, 2073–2083. [CrossRef] [PubMed]
94. Xi, Y.; Yuefen, P.; Wei, W.; Quan, Q.; Jing, Z.; Jiamin, X.; Shuwen, H. Analysis of prognosis, genome, microbiome, and microbial metabolome in different sites of colorectal cancer. *J. Transl. Med.* **2019**, *17*, 1–22. [CrossRef]
95. Kamphuis, J.; Mercier-Bonin, M.; Eutamène, H.; Théodorou, V. Mucus organisation is shaped by colonic content; a new view. *Sci. Rep.* **2017**, *7*, 1–13. [CrossRef]
96. Paone, P.; Cani, P.D. Mucus barrier, mucins and gut microbiota: The expected slimy partners? *Gut* **2020**, *69*, 2232–2243. [CrossRef]

97. Luu, T.H.; Michel, C.; Bard, J.-M.; Dravet, F.; Nazih, H.; Bobin-Dubigeon, C. Intestinal Proportion ofBlautiasp. is Associated with Clinical Stage and Histoprognostic Grade in Patients with Early-Stage Breast Cancer. *Nutr. Cancer* **2017**, *69*, 267–275. [CrossRef]
98. Wu, A.H.; Tseng, C.; Vigen, C.; Yu, Y.; Cozen, W.; Garcia, A.A.; Spicer, D. Gut microbiome associations with breast cancer risk factors and tumor characteristics: A pilot study. *Breast Cancer Res. Treat.* **2020**, *182*, 451–463. [CrossRef]
99. Zhuang, H.; Cheng, L.; Wang, Y.; Zhang, Y.-K.; Zhao, M.-F.; Liang, G.-D.; Zhang, M.-C.; Li, Y.-G.; Zhao, J.-B.; Gao, Y.-N.; et al. Dysbiosis of the Gut Microbiome in Lung Cancer. *Front. Cell. Infect. Microbiol.* **2019**, *9*, 112. [CrossRef]
100. Budinska, E.; Popovici, V.; Tejpar, S.; D'Ario, G.; Lapique, N.; Sikora, K.O.; Di Narzo, A.F.; Yan, P.; Hodgson, J.G.; Weinrich, S.; et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **2013**, *231*, 63–76. [CrossRef]
101. He, Z.; Gharaibeh, R.Z.; Newsome, R.C.; Pope, J.L.; Dougherty, M.; Tomkovich, S.; Pons, B.; Mirey, G.; Vignard, J.; Hendrixson, D.R.; et al. Campylobacter jejuni promotes colorectal tumorigenesis through the action of cytolethal distending toxin. *Gut* **2019**, *68*, 289–300. [CrossRef] [PubMed]
102. Drewes, J.L.; White, J.R.; Dejea, C.M.; Fathi, P.; Iyadorai, T.; Vadivelu, J.; Roslani, A.C.; Wick, E.C.; Mongodin, E.F.; Loke, M.F.; et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *Npj Biofilms Microbiomes* **2017**, *3*, 1–12. [CrossRef] [PubMed]
103. Kaplan, C.W.; Lux, R.; Haake, S.K.; Shi, W. TheFusobacterium nucleatumouter membrane protein RadD is an arginine-inhibitable adhesin required for inter-species adherence and the structured architecture of multispecies biofilm. *Mol. Microbiol.* **2009**, *71*, 35–47. [CrossRef]
104. Tomkovich, S.; Dejea, C.M.; Winglee, K.; Drewes, J.L.; Chung, L.; Housseau, F.; Pope, J.L.; Gauthier, J.; Sun, X.; Mühlbauer, M.; et al. Human colon mucosal biofilms from healthy or colon cancer hosts are carcinogenic. *J. Clin. Investig.* **2019**, *129*, 1699–1712. [CrossRef] [PubMed]
105. Jorth, P.; Turner, K.H.; Gumus, P.; Nizam, N.; Buduneli, N.; Whiteley, M. Metatranscriptomics of the Human Oral Microbiome during Health and Disease. *mBio* **2014**, *5*, e01012-14. [CrossRef] [PubMed]
106. Liu, Y.; Geng, R.; Liu, L.; Jin, X.; Yan, W.; Zhao, F.; Wang, S.; Guo, X.; Ghimire, G.; Wei, Y. Gut Microbiota-Based Algorithms in the Prediction of Metachronous Adenoma in Colorectal Cancer Patients Following Surgery. *Front. Microbiol.* **2020**, *11*, 1106. [CrossRef] [PubMed]
107. Alexander, J.L.; Wilson, I.D.; Teare, J.; Marchesi, J.; Nicholson, J.K.; Kinross, J. Gut microbiota modulation of chemotherapy efficacy and toxicity. *Nat. Rev. Gastroenterol. Hepatol.* **2017**, *14*, 356–365. [CrossRef] [PubMed]
108. Chew, S.-S.; Tan, L.T.-H.; Law, J.W.-F.; Pusparajah, P.; Goh, B.-H.; Ab Mutalib, N.S.; Lee, L.-H. Targeting Gut Microbial Biofilms—A Key to Hinder Colon Carcinogenesis? *Cancers* **2020**, *12*, 2272. [CrossRef] [PubMed]
109. Kim, G.W.; Kim, Y.-S.; Lee, S.H.; Park, S.G.; Kim, D.H.; Cho, J.Y.; Hahm, K.B.; Hong, S.P.; Yoo, J.-H. Periodontitis is associated with an increased risk for proximal colorectal neoplasms. *Sci. Rep.* **2019**, *9*, 7528. [CrossRef]
110. Horz, H.-P.; Scheer, S.; Huenger, F.; Vianna, M.E.; Conrads, G. Selective isolation of bacterial DNA from human clinical specimens. *J. Microbiol. Methods* **2008**, *72*, 98–102. [CrossRef]
111. Walker, S.P.; Tangney, M.; Claesson, M. Sequence-Based Characterization of Intratumoral Bacteria—A Guide to Best Practice. *Front. Oncol.* **2020**, *10*, 179. [CrossRef]
112. Heravi, F.S.; Zakrzewski, M.; Vickery, K.; Hu, H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. *J. Microbiol. Methods* **2020**, *170*, 105856. [CrossRef]

[***18***] Stenvang J, **Budinská E,** van Cutsem E, Bosman F, Popovici V, Brünner N. An Explorative Analysis of *ABCG2/TOP-1* mRNA Expression as a Biomarker Test for FOLFIRI Treatment in Stage III Colon Cancer Patients: Results from Retrospective Analyses of the PETACC-3 Trial. Cancers (Basel). 2020 Apr 15;12(4):977. doi: 10.3390/cancers12040977. PMID: 32326511; PMCID: PMC7226226.

cancers

MDPI

Article

An Explorative Analysis of *ABCG2/TOP-1* mRNA Expression as a Biomarker Test for FOLFIRI Treatment in Stage III Colon Cancer Patients: Results from Retrospective Analyses of the PETACC-3 Trial

Jan Stenvang [1,2], Eva Budinská [3], Eric van Cutsem [4], Fred Bosman [5], Vlad Popovici [3,*,†] and Nils Brünner [1,2,*,†]

[1] Section of Molecular Disease Biology, Institute of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark; stenvang@sund.ku.dk
[2] Scandion Oncology, Symbion, 2100 Copenhagen, Denmark
[3] RECETOX, Faculty of Science, Masarykova Univerzita, 625 00 Brno, Czech Republic; budinska@recetox.muni.cz
[4] Digestive Oncology Department, University Hospitals Leuven and KU Leuven, 3001 Leuven, Belgium; eric.vancutsem@uzleuven.be
[5] University Institute of Pathology, University of Lausanne, 1011 Lausanne, Switzerland; Fred.Bosman@chuv.ch
* Correspondence: popovici@recetox.muni.cz (V.P.); nb@scandiononcology.com (N.B.)
† These authors contributed equally to the study.

Received: 24 March 2020; Accepted: 8 April 2020; Published: 15 April 2020

Abstract: Biomarker-guided treatment for patients with colon cancer is needed. We tested ABCG2 and topoisomerase 1 (TOP1) mRNA expression as predictive biomarkers for irinotecan benefit in the PETACC-3 patient cohort. The present study included 580 patients with mRNA expression data from Stage III colon cancer samples from the PETACC-3 study, which randomized the patients to Fluorouracil/leucovorin (5FUL) +/− irinotecan. The primary end-points were recurrence free survival (RFS) and overall survival (OS). Patients were divided into one group with high ABCG2 expression (above median) and low TOP-1 expression (below 75 percentile) ("resistant") (n = 216) and another group including all other combinations of these two genes ("sensitive") (n = 364). The rationale for the cut-offs were based on the distribution of expression levels in the PETACC-3 Stage II set of patients, where ABCG2 was unimodal and TOP1 was bimodal with a high expression level mode in the top quarter of the patients. Cox proportional hazards regression was used to estimate the hazard ratios and the association between variables and end-points and log-rank tests to assess the statistical significance of differences in survival between groups. Kaplan-Meier estimates of the survival functions were used for visualization and estimation of survival rates at specific time points. Significant differences were found for both RFS (Hazard ratio (HR): 0.63 (0.44–0.92); p = 0.016) and OS (HR: 0.60 (0.39–0.93); p = 0.02) between the two biomarker groups when the patients received FOLFIRI (5FUL+irinotecan). Considering only the Microsatellite Stable (MSS) and Microsatellite Instability-Low (MSI-L) patients (n = 470), the differences were even more pronounced. In contrast, no significant differences were observed between the groups when patients received 5FUL alone. This study shows that the combination of ABCG2 and TOP1 gene expression significantly divided the Stage III colon cancer patients into two groups regarding benefit from adjuvant treatment with FOLFIRI but not 5FUL.

Keywords: biomarkers; ABCG2; TOP-1; adjuvant irinotecan; colon cancer

1. Introduction

Based on the results from the MOSAIC prospective randomized clinical trial (PRCT) [1], treatment of high risk Stage II and of Stage III colon cancer (CC) patients currently consists of 5-Fluorouracil or Xeloda plus leucovorin (5FUL) plus oxaliplatin (FOLFOX or XELOX).

Presently, irinotecan is not used in adjuvant treatment of primary CC but only in the metastatic setting [2]. The reason for this is that none of two high-powered and independent PRCTs (PETACC-3 [3] and CALGB 89803 [4]), including high risk Stage II and Stage III colon cancer and randomizing patients to 5FUL ± irinotecan, could demonstrate a significant difference between the treatment groups with respect to recurrence free survival (RFS) or overall survival (OS).

With a five-year recurrence rate of approximately 30% following adjuvant FOLFOX/XELOX treatment of Stage III CC patients [1], there is obviously a need for other adjuvant treatment modalities using drugs with different mechanisms of action than FOLFOX/XELOX. These treatments must be accompanied by predictive biomarkers, allowing rational allocation of individual patients to the most effective regimen.

Although PRCTs have not shown any added benefit from adjuvant irinotecan of CC cancer when co-administered with 5FUL, it is conceivable that specific subgroups of patients may benefit from the addition of irinotecan treatment but that these patients are concealed within the total patient population. Moreover, given the fundamental differences in molecular mechanisms of action of oxaliplatin and irinotecan as well as the different molecular mechanisms underlying resistance to these two drugs [5], it is likely that the groups of patients benefitting from adjuvant FOLFOX/XELOX or FOLFIRI treatments, respectively, are only partly overlapping, if at all.

We recently characterized isogeneic pairs of CRC cell lines selected for resistance to SN38 (the active metabolite of irinotecan) or oxaliplatin [5]. We identified several genetic aberrations associated with SN38 resistance; in particular, the xenobiotic drug transporter ABCG2 was found to be the most up-regulated gene in the SN38 resistant cell lines [5]. Subsequent functional analyses of this gene demonstrated its major role in SN38 resistance [5]. Moreover, downregulation of the irinotecan target, the topoisomerase-1 enzyme (Top-1), has also been observed in our SN38 resistant cancer cells [6]. Of specific interest is that the resistance mechanisms in our three oxaliplatin-resistant colorectal cancer cell lines [5] did not include regulation of ABCG2 or TOP1 mRNA. In recent publications [5,7], we correlated each of ABCG2 and TOP1 mRNA expression to patient outcome in a subset of Stage III colon cancer patients enrolled in the PETACC-3 study. A trend was demonstrated towards high ABCG2 mRNA expression being correlated with shorter recurrence-free survival (RFS) and shorter overall survival (OS) when compared to patients with low ABCG2 mRNA expression [5], and high TOP1 expression was significantly associated with longer OS but not RFS in FOLFIRI treated patients [7]. We now hypothesize that low TOP1 and high ABCG2 expression (“resistant patients”) define patients who will not benefit from irinotecan containing adjuvant chemotherapy while any other combination of these two genes defines patients (“sensitive patients”) who will benefit from the addition of irinotecan to 5FUL. The present study was designed to test this hypothesis.

2. Results

2.1. Patient Characteristics

For a detailed description, including a CONSORT diagram on the selection of the present PETACC-3 cohort, please see [7]. Table 1 shows the clinicopathological characteristics of the $n = 580$ stage III CC patients included in the study. For comparison, the clinicopathological characteristics of the full set of 2315 patients from the PETACC-3 Stage III CC patient cohort are included. With gender composition as exception (the subpopulation is slightly enriched in males), the present study population was representative of the global PETACC-3 study population.

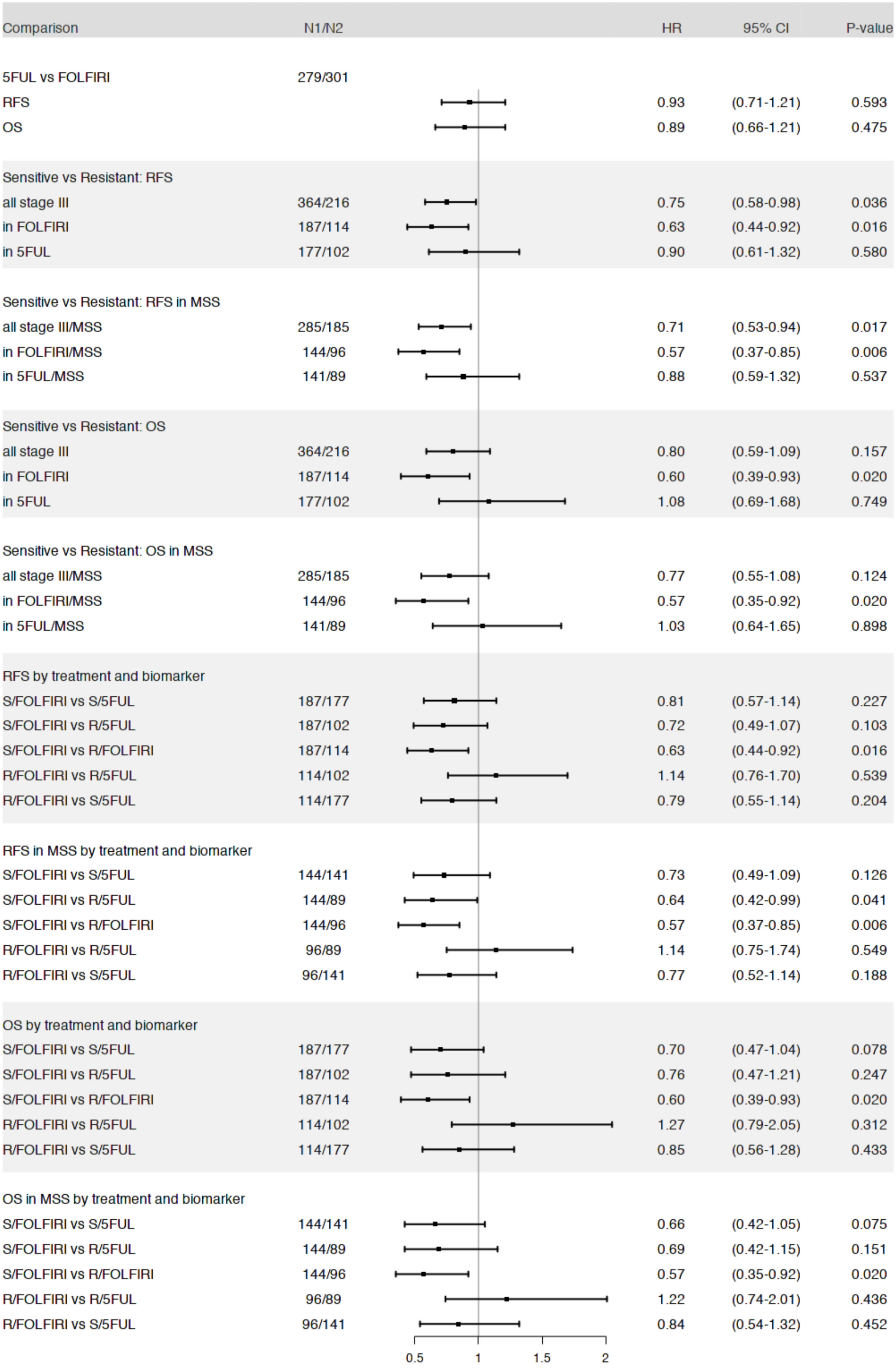
Table 1. Population characteristics for the whole PETACC-3/Stage III and the study subpopulation. The only statistically significant difference was between male/female proportions (* starred covariate in the table; $p = 0.025$). The missing values (denoted NA (not available)) were not considered when computing the proportions. Microsatellite Instability (MSI) Status is divided into MSI High (MSI-H), MSI Low (MSI-L) and Microsatellite Stable (MSS).

Variables	All PETACC-3 Stage III (n = 2315)	Study Subpopulation (n = 580)
Age (mean (sd))	58.35 (10.54)	58.86 (10.44)
Sex * (n (%))		
Male	1263 (54.6)	347 (59.8)
Female	1052 (45.4)	233 (40.2)
Treatment (n (%))		
5FUL	1157 (50.0)	279 (48.1)
FOLFIRI	1158 (50.0)	301 (51.9)
Site (n (%))		
left	1422 (61.4)	366 (63.1)
right	893 (38.6)	214 (36.9)
Grade (n (%))		
1,2	877 (88.1)	512 (88.9)
3,4	119 (11.9)	64 (11.1)
NA	1319	4
T-stage (n (%))		
T1, T2	196 (8.5)	51 (8.8)
T3	1766 (76.4)	438 (75.5)
T4	351 (15.2)	91 (15.7)
NA	2	0
N-stage (n (%))		
N0, N1	1496 (64.4)	377 (65.0)
N2	819 (35.4)	203 (35.0)
Mucinous histology (n (%))		
No	807 (81.0)	477 (82.8)
Yes	189 (19.0)	99 (17.2)
NA	1319	4
MSI status (n (%))		
MSI-H	106 (12.1)	51 (9.8)
MSI-L, MSS	772 (87.9)	470 (90.2)
NA	1437	59
BRAF V600E (n (%))		
mutated	78 (8.4)	37 (6.7)
wild type	848 (91.6)	512 (93.3)
NA		31
KRAS codon 12, 13 (n (%))		
mutated	364 (39.6)	219 (40.1)
wild type	556 (60.4)	327 (59.9)
NA	1395	34

Table 2 shows the correlations between clinicopathological parameters and ABCG2 gene expression and TOP1 gene expression, respectively. TOP1 was associated with site, grade, mucinous histology and KRAS mutational status, while ABCG2 was associated with site, MSI and BRAF mutational status. These observations suggest a possible association of ABCG2 with BRAF mutated pathway, and of TOP1 with BRAF-mutant-like (t-test p-value < 0.001) [8]. In univariate analysis including all 580 patients stratified by treatment arm, no significant benefit from irinotecan addition was found either for RFS or for OS (Figure 1) and thus the selected subgroup does not differ from the main PETACC-3 population with regard to treatment effect.

Table 2. Comparison of expression levels between various stratifications in the study subpopulation for ABCG2 and TOP1 genes, respectively. For each gene, the mean and standard deviation of the expression levels (log2) are indicated and the corresponding p-values from Student's *t*-test (significant are emphasized by italic) for binary categories and ANOVA for multiple categories.

Stratification Factor	n (%)	TOP1 (Mean (sd))	ABCG2 (Mean (sd))
Site			
left	366 (63.1)	4.85 (1.00)	2.43 (0.55)
right	214 (36.9)	4.59 (0.98)	2.55 (0.73)
p-value		*0.002*	*0.025*
Grade			
1, 2	512 (88.9)	4.79 (0.97)	2.46(0.59)
3,4	64 (11.1)	4.49 (1.21)	2.59 (0.85)
p-value		*0.026*	0.120
T-stage			
T1, T2	51 (8.8)	4.86 (1.10)	2.47 (0.46)
T3	438 (75.5)	4.75 (0.96)	2.48 (0.65)
T4	91 (15.7)	4.74 (1.12)	2.46 (0.55)
p-value		0.725	0.961
N-stage			
N0, N1	377 (65.0)	4.80 (0.95)	2.44 (0.52)
N2	203 (35.0)	4.68 (1.08)	2.54 (0.78)
p-value		0.162	0.079
Mucinous histology			
no	477 (82.8)	4.85 (0.98)	2.48 (0.63)
yes	99 (17.2)	4.32 (0.97)	2.43 (0.61)
p-value		< *0.001*	0.411
MSI status			
MSI-H	51 (9.8)	4.52 (1.01)	2.20 (0.40)
MSI-L, MSS	470 (90.2)	4.78 (0.99)	2.49 (0.59)
p-value		0.074	*0.001*
BRAF V600E mutation			
mutated	37 (6.7)	4.55 (1.15)	2.81 (0.98)
wild type	512 (93.3)	4.77 (0.99)	2.44 (0.53)
p-value		0.202	< *0.001*
KRAS codon 12, 13			
mutated	219 (40.1)	4.65 (0.97)	2.45 (0.55)
wild type	327 (59.9)	4.82 (1.01)	2.48 (0.59)
p-value		*0.040*	0.549



Comparison	N1/N2	HR	95% CI	P-value
5FUL vs FOLFIRI	279/301			
RFS		0.93	(0.71-1.21)	0.593
OS		0.89	(0.66-1.21)	0.475
Sensitive vs Resistant: RFS				
all stage III	364/216	0.75	(0.58-0.98)	0.036
in FOLFIRI	187/114	0.63	(0.44-0.92)	0.016
in 5FUL	177/102	0.90	(0.61-1.32)	0.580
Sensitive vs Resistant: RFS in MSS				
all stage III/MSS	285/185	0.71	(0.53-0.94)	0.017
in FOLFIRI/MSS	144/96	0.57	(0.37-0.85)	0.006
in 5FUL/MSS	141/89	0.88	(0.59-1.32)	0.537
Sensitive vs Resistant: OS				
all stage III	364/216	0.80	(0.59-1.09)	0.157
in FOLFIRI	187/114	0.60	(0.39-0.93)	0.020
in 5FUL	177/102	1.08	(0.69-1.68)	0.749
Sensitive vs Resistant: OS in MSS				
all stage III/MSS	285/185	0.77	(0.55-1.08)	0.124
in FOLFIRI/MSS	144/96	0.57	(0.35-0.92)	0.020
in 5FUL/MSS	141/89	1.03	(0.64-1.65)	0.898
RFS by treatment and biomarker				
S/FOLFIRI vs S/5FUL	187/177	0.81	(0.57-1.14)	0.227
S/FOLFIRI vs R/5FUL	187/102	0.72	(0.49-1.07)	0.103
S/FOLFIRI vs R/FOLFIRI	187/114	0.63	(0.44-0.92)	0.016
R/FOLFIRI vs R/5FUL	114/102	1.14	(0.76-1.70)	0.539
R/FOLFIRI vs S/5FUL	114/177	0.79	(0.55-1.14)	0.204
RFS in MSS by treatment and biomarker				
S/FOLFIRI vs S/5FUL	144/141	0.73	(0.49-1.09)	0.126
S/FOLFIRI vs R/5FUL	144/89	0.64	(0.42-0.99)	0.041
S/FOLFIRI vs R/FOLFIRI	144/96	0.57	(0.37-0.85)	0.006
R/FOLFIRI vs R/5FUL	96/89	1.14	(0.75-1.74)	0.549
R/FOLFIRI vs S/5FUL	96/141	0.77	(0.52-1.14)	0.188
OS by treatment and biomarker				
S/FOLFIRI vs S/5FUL	187/177	0.70	(0.47-1.04)	0.078
S/FOLFIRI vs R/5FUL	187/102	0.76	(0.47-1.21)	0.247
S/FOLFIRI vs R/FOLFIRI	187/114	0.60	(0.39-0.93)	0.020
R/FOLFIRI vs R/5FUL	114/102	1.27	(0.79-2.05)	0.312
R/FOLFIRI vs S/5FUL	114/177	0.85	(0.56-1.28)	0.433
OS in MSS by treatment and biomarker				
S/FOLFIRI vs S/5FUL	144/141	0.66	(0.42-1.05)	0.075
S/FOLFIRI vs R/5FUL	144/89	0.69	(0.42-1.15)	0.151
S/FOLFIRI vs R/FOLFIRI	144/96	0.57	(0.35-0.92)	0.020
R/FOLFIRI vs R/5FUL	96/89	1.22	(0.74-2.01)	0.436
R/FOLFIRI vs S/5FUL	96/141	0.84	(0.54-1.32)	0.452

Figure 1. Summary of main survival analysis results: log-rank tests were used to assess the statistical significance of survival differences between groups of interest. Each section of results refers to a set of related tests. The differences assessed are given in the first column, the corresponding sample sizes in the second column (N1/N2: sample size of the first and second group, respectively), while the columns 3–6 summarize the test results in terms of hazard ratios and 95% confidence intervals (plots in column 3) and corresponding *p*-values (log-rank test—column 6). In the first column, "sensitive" was abbreviated as "S" and "resistant" as "R", respectively. Thus, "S/FOLFIRI" stands for "sensitive under FOLFIRI treatment" etc.

2.2. Combining TOP1 and ABCG2 mRNA Expression

The Spearman's correlation coefficient between ABCG2 and TOP1 gene expression was r = 0.046 (Figure S1). There were 216 patients in the ABCG2 high/TOP1 low ("resistant patients") and 364 patients in the "sensitive patient" group. When stratifying the whole set of patients (n = 580) according to ABCG2/TOP1 status, a significantly better RFS (Hazard Ratio (HR): 0.75; 95% confidence interval CI: 0.58–0.98; p = 0.036) was observed in the "sensitive patient" group as compared to the ABCG2 high/TOP1 low "resistant patient" group (Figure 1 and Figure S2; online only). When stratifying each of the two treatment groups according to the proposed test, the separation between the "sensitive" and "resistant" patient groups in terms of RFS was significant in the FOLFIRI arm (HR: 0.63; 95% CI: 0.44–0.92; p = 0.016) but not in the 5FUL arm (Figure 1, Figure 2A,B).

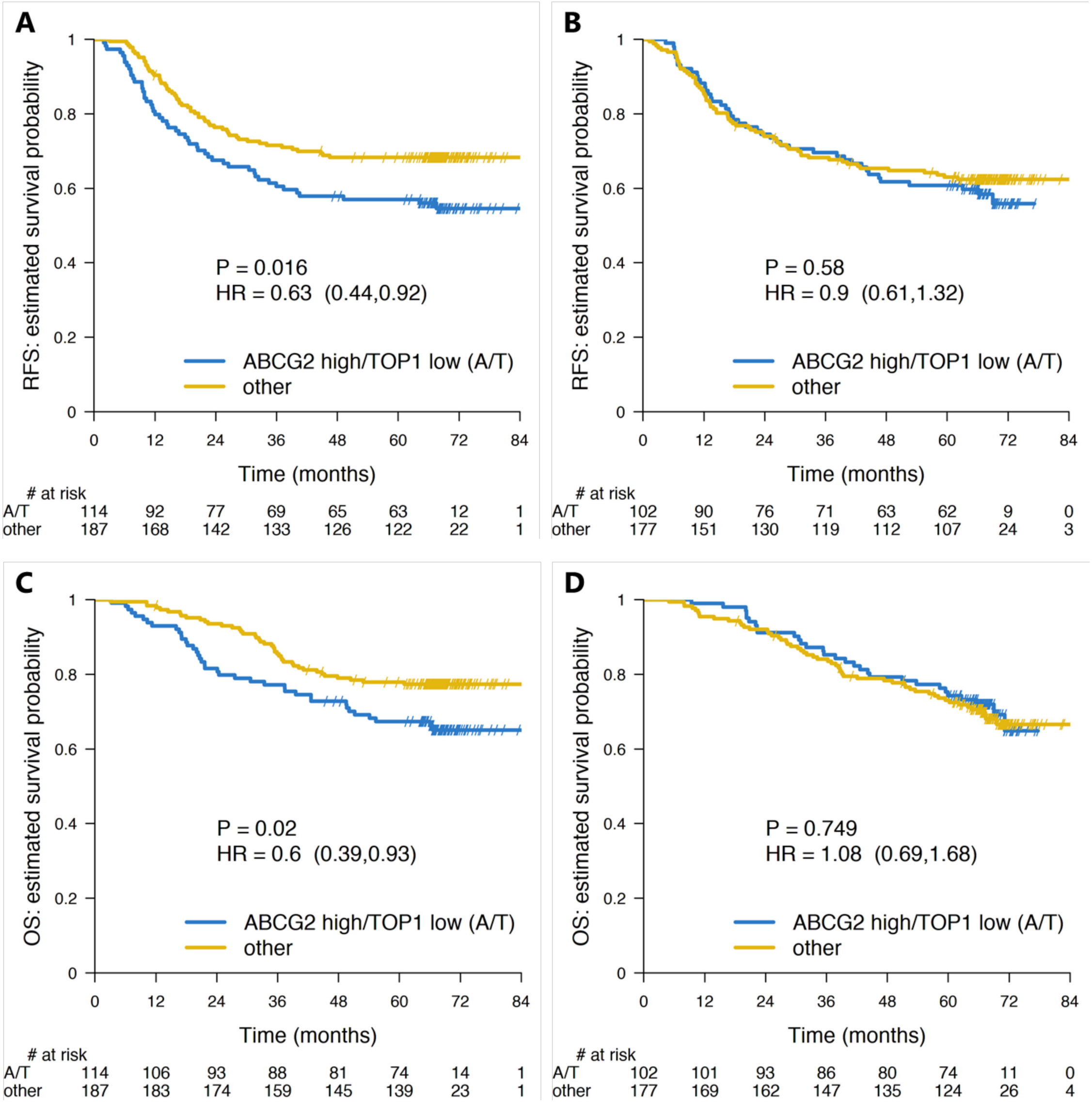


Figure 2. Survival plots (Kaplan-Meier estimates) for "resistant" (ABCG2-high/TOP1-low, abbreviated A/T under the plots) and "sensitive" (all other combinations of ABCG2 and TOP1 genes) patient groups in whole Stage III cohort (n = 580). The four plots show the RFS of "resistant" (blue line) and "sensitive" (gold line) under (**A**) Fluorouracil/leucovorin (5FUL) + irinotecan (FOLFIRI) and (**B**) 5FUL treatments and the overall survival (OS) of the same groups under (**C**) FOLFIRI and (**D**) 5FUL treatments, respectively. Numbers at risk are given under each plot.

In terms of relative 3- and 5-years RFS, the patients in the "sensitive" group performed better only under FOLFIRI treatment (relative benefit of 18.2% and 19.9% at 3- and 5-years, respectively–Table 3). The complete pairwise comparisons between all combinations of test group ("sensitive" vs. "resistant" patients) and treatment arm (FOLFIRI vs. 5FUL) (six comparisons) did not yield any statistically significant difference, aside from that between "sensitive" and "resistant" patients groups within the FOLFIRI arm (Figure 1 and Figure S3 (online only)).

Table 3. Summary of patient survival rates by treatment and biomarker group (R: resistant, S: sensitive) at 3 and 5 years, respectively. The relative benefit is denoted by (S−R)/R.

				FOLFIRI				
End-Point	**S vs. R**		**3-Year Survival Rates**			**5-Year Survival Rates**		
	HR (95% CI)	***p*-Value**	**S (%) (95% CI)**	**R (%) (95% CI)**	**(S−R)/R (%)**	**S (%) (95% CI)**	**R (%) (95% CI)**	**(S−R)/R (%)**
RFS	0.63 (0.44–0.92)	*0.016*	71.5 (65.3–78.3)	60.5 (52.2–70.5)	18.2	68.3 (61.9–75.3)	57.0 (48.6–66.8)	19.9
OS	0.60 (0.39–0.93)	*0.020*	85.5 (80.6–90.7)	77.2 (69.9–85.3)	10.8	77.9 (72.2–84.1)	67.3 (59.2–76.6)	15.7
				5FUL				
	S vs R		**3-year survival rates**			**5-year survival rates**		
	HR (95% CI)	***p*-value**	**S (%) (95% CI)**	**R (%) (95% CI)**	**(S−R)/R (%)**	**S (%) (95% CI)**	**R (%) (95% CI)**	**(S−R)/R (%)**
RFS	0.90 (0.61–1.32)	0.58	68.3 (61.7–75.5)	69.6 (61.2–79.1)	−0.02	63.0 (56.3–70.6)	60.8 (52.0–71.0)	0.04
OS	1.08 (0.69–1.68)	0.75	84.1 (78.8–89.7)	85.2 (78.6–92.4)	−0.01	73.1 (66.8–80.0)	74.3 (66.2–83.3)	−0.02

Abbreviations: HR (Hazard Ratio), CI (Confidence Interval), FOLFIRI (Fluorouracil/leucovorin (5FUL) + irinotecan), RFS (Recurrence Free Survival), OS (Overall Survival).

We also pooled all the 5FUL-only treated patients and estimated the 3- and 5-year RFS (3-years RFS: 68.8%; 5-year RFS: 62.2%). The relative benefit in 3-year and 5-year RFS between FOLFIRI-treated "sensitive" patients and all 5FUL alone treated patients were 4.1% and 9.9%, respectively, in favor of FOLFIRI.

Similar analyses were performed for OS as endpoint. In the whole Stage III population, no statistically significant difference was found between "sensitive" and "resistant" patients (Figure 1 and Figure S4 (online only)). However, when analyzing by treatment arm, the FOLFIRI-treated patients labeled as "sensitive" by the test had a significantly longer OS (HR: 0.6; 95% CI: 0.35–0.92; $p = 0.02$) (Figures 1 and 2C), while no such difference could be detected in 5FUL only treated patients (Figures 1 and 2D). When combining the 5FUL patients into one group and then comparing this pooled group with each of the two FOLFIRI treated groups, no significant differences in OS were observed (Figure S5; online only). Nevertheless, the "sensitive" patients treated with FOLIFIRI seemed to fare better with 3- and 5-year relative gains of 1.2% and 6%, respectively (Figure S5; online only). The pairwise comparisons of all possible combinations between test group and treatment arm did not reveal any significant difference with the exception of the one between "sensitive" and "resistant" patients treated with FOLFIRI, already discussed above (Figure 1).

2.3. ABCG2 and TOP1 in MSS Plus MSI-L Patient Subgroup

Due to the low number of Microsatellite Instable (MSI) tumors, we focused our analyses on the Stage III Microsatellite Stable (MSS) plus Microsatellite Instable-Low (MSI-L) tumors ($n = 470$). MSS and MSI-L patients in "sensitive patients" treated with FOLFIRI had a significant better RFS (HR: 0.57; 95% CI: 0.37–0.85; $p = 0.006$) (Figure 3A) and OS (HR: 0.57, 95% CI: 0.35–0.92; $p = 0.02$) (Figure 3B) than patients in the "resistant" group. Stratifying the 5FUL only treated MSS patients by the ABCG2 and TOP1 test did not result in any significant separation of the patients for RFS or OS (Figure 1). The 5-year RFS for all MSS plus MSI-L patients treated with 5FUL alone was 59.7% (95% CI: 53.7–66.5), while for "sensitive patients" treated with FOLFIRI it was 69.3% (95% CI: 62.1–77.3), resulting in a relative gain of 15.9% in favor of the latter. When also dichotomizing the 5FUL-only treated group with the biomarker test and when comparing to the FOLFIRI arm, it was seen that FOLFIRI treatment of "sensitive patients" resulted in a 7.3% and 14.3% relative gain in 3-year and 5-year RFS in comparison with the equivalent group treated with 5FUL alone, however without reaching statistical significance. When considering all possible pairwise comparisons of groups defined by the test and treatment arm (Figures S6–S8 (online only) for RFS and OS, respectively) the only significant differences were between "sensitive" and "resistant" patients treated with FOLFIRI and between FOLFIRI-treated "resistant patients" and 5FUL-treated group 1 patients (Figure 1).

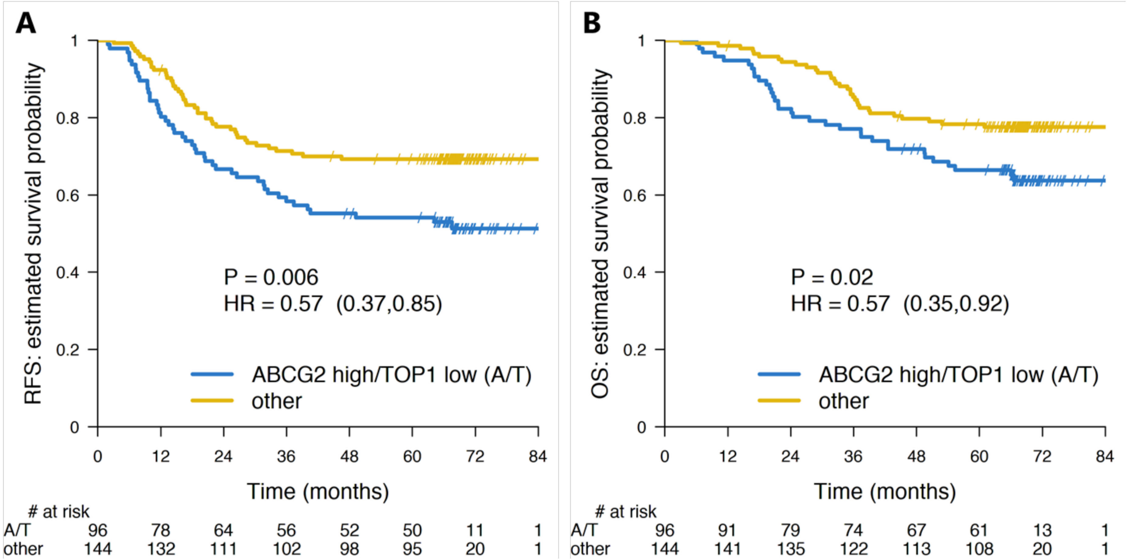


Figure 3. Survival plots (Kaplan-Meier estimates) for "resistant" (ABCG2-high/TOP1-low, abbreviated A/T under the plots) and "sensitive" (all other combinations of ABCG2 and TOP1 genes) patient groups in stage III/MSS (Microsatellite Stable) subset (n = 470). The two plots show the (**A**) Recurrence free survival (RFS) and (**B**) Overall Survival (OS) of "resistant" (blue line) and "sensitive" (gold line) under FOLFIRI treatment, respectively. Numbers at risk are given under each plot.

2.4. ABCG2/TOP1 Status as Independent Predictor in Multivariable Models

We tested the independence of ABCG2/TOP1 status in multivariable models including tumor site, MSI status, mucinous histology, and BRAF and KRAS mutation status (without interaction terms). In least absolute shrinkage and selection operator (LASSO) [9] penalized regression analyses, ABCG2/TOP1 status was selected as the most important variable for RFS both in the whole population and in the FOLFIRI-treated arm. In 5FUL, none of the tested variables was found to be significant. Similar results were obtained in the MSS subpopulation, with ABCG2/TOP1 status being selected as the most important variable in whole MSS and in MSS FOLFIRI-treated subpopulations, but not in MSS 5FUL.

In multivariable Cox regression, after including all the variables selected by penalized regression, ABCG2/TOP1 status had a corresponding adjusted HR: 0.75 (95% CI: 0.57–1.00, p = 0.052) for the whole population and HR: 0.72 (95% CI: 0.53–0.96, p = 0.028) for the MSS subpopulation. In FOLFIRI arm, ABCG2/TOP1 status had a corresponding HR: 0.63 (95% CI: 0.42–0.94, p = 0.020) for all patients and HR: 0.57 (95% CI: 0.38–0.87, p = 0.009) for the MSS subpopulation. See Supplemental Materials—Multivariable Regression section.

3. Discussion

Irinotecan is a topoisomerase 1 poison and by binding to the Top1 enzyme, toxic complexes are formed leading to induction of apoptosis. We therefore hypothesized that a higher Top1 level in cancer cells is associated with more toxic effects of irinotecan. ABCG2 is a xenobiotic drug efflux pump being involved in outwards transportation of SN38 from cells. An additional hypothesis therefore is that a high cellular level of ABCG2 is associated with less cytotoxic effects of irinotecan.

In a previous study, which included the 580 PETACC-3 patients [5], we found in Stage III CC patients a trend towards association between high ABCG2 expression and poor patient outcome in the irinotecan containing treatment group, but not in the 5FUL only treated group. In another retrospective PETACC-3 study [7], we reported for low TOP1 expression a trend towards an association with short RFS and a borderline significant association with shorter OS in the irinotecan treated patients but not in the 5FUL treated patients [7]. On the assumption that more than one molecular resistance mechanism is involved in irinotecan resistance [5], we now combined ABCG2 and TOP1 expression status in a single dichotomous parameter and hypothesized that patients with a tumor with a high ABCG2 and a low TOP1 expression level might represent those with a low response to irinotecan added to adjuvant 5FUL treatment. As presented in Figure 2, our data confirm this hypothesis in showing that ABCG2/TOP1 status is significantly associated with RFS and OS in Stage III CC patients receiving adjuvant irinotecan containing chemotherapy. In contrast, ABCG2/TOP1 status was neither associated with RFS nor with OS in patients receiving 5FUL only as adjuvant treatment, which is consistent with a predictive rather than a prognostic value.

We compared our results with those published from the MOSAIC study [1], in which 2216 stage III CC patients were randomly assigned to receive 5FUL alone or in combination with oxaliplatin (FOLFOX) for six months. A significant difference ($p = 0.003$) in disease-free survival (DFS) in favor of FOLFOX with a 5-year 8.8% relative increase in DFS in the FOLFOX treated patients was noted and adjuvant FOLFOX is now the standard of care in Stage III CC patients. When we compared 5-year RFS between FOLFIRI "sensitive patients" and the total 5FUL only treated group in our study, we noted that RFS in FOLFIRI treated FOLFIRI "sensitive patients" was 9.9% higher than that of all patients treated with 5FUL only. Thus, the benefit from adjuvant systemic treatment in the MOSAIC study and in the PETACC-3 subgroup of patients with FOLFIRI "sensitive" tumors were comparable at 5 years (the only difference between DFS in the MOSAIC study and RFS in the PETACC-3 study was the inclusion of a second malignancy in the RFS).

We also performed subgroup analyses, including only MSS + MSI-L patients, which led to even more significant results than those reported for the whole cohort (Figure 3). Klingbiel et al. [10] previously reported that in the PETACC-3 study, MSI status had no effect on survival of FOLFIRI treated patients, neither on RFS nor on OS. Moreover, an interaction test between treatment and MSI status in Stage III patients was not significant. When we included only MSS+MSI-L patients, the differences in RFS and OS between FOLFIRI and 5FUL patients and ABCG2/TOP1 status became more pronounced but still did not reach statistical significance, most probably due to the low number of included patients.

An interesting point is that the function of ABCG2 can be inhibited in patients [11]. In Scandion Oncology, we develop novel drugs to inhibit ABCG2 [12,13]. When these drugs have been tested in regular clinical phase II trials in patients with metastatic and irinotecan resistant colorectal cancer, they can be taken into randomized clinical testing including Stage III colon cancer patients with high ABCG2 expression. We recently analytically validated commercial antibodies for immunohistochemical (IHC) staining of ABCG2 on formalin-fixed formalin embedded colorectal cancer tissue and identified the BXP21 antibody to fulfill requirements for ABCG2 IHC [14].

The major strength of our study lies in the design. In the PETACC-3 PRCT, 5FUL treatment constituted the backbone and irinotecan was added to half of the patients only. This design lends itself to study biomarkers predictive of irinotecan response and to separate a potential predictive from a prognostic impact [15]. Moreover, RFS and OS are valid endpoints for estimating the effect of predictive biomarkers. Finally, the choice of ABCG2 and TOP1 as potential biomarkers for irinotecan sensitivity/resistance was based on a hypothesis derived from results of our in vitro studies on cell lines [5,6].

The weakness of the study is related to the lack of an independent validation cohort. However, only one other PRCT (CALGB 89803) has investigated the impact of adding irinotecan to 5FUL in the adjuvant treatment of Stage III CC [4]. Unfortunately, no mRNA expression data are available from the CALGB 89803 study.

4. Materials and Methods

4.1. Patients

The set of patients considered for the present study consisted of all Stage III patients with good quality mRNA expression data (n = 580) from the PETACC-3 study [3]. For further information on patient characteristics, inclusion and exclusion criteria, treatment schedules and follow-up, please see the original publication [3]. n = 279 of these patients had been randomized to receive adjuvant 5FUL only and n = 301 patients received irinotecan in addition to 5FUL. Further details on the present study population are given in Table 1.

All patients signed an informed consent form, allowing collection of tumor tissue for future translational research. Approval for the present translational study was obtained from the PETACC-3 Translational Research Working Party.

4.2. Gene Expression Analyses

As previously described [16], total RNA was extracted from formalin fixed paraffin embedded (FFPE) blocks of the primary cancers. The RNA was amplified and hybridized to the Almac Colorectal Cancer DSA microarray platform (Almac, Craigavon, UK). Whole-genome gene expression data is publicly available from ArrayExpress under accession number E-MTAB-990 [16].

4.3. Statistical Methods

The present study was prospective-retrospective in nature. The statistical plan and the applied cutoff values were defined prior to the study. We used the original PETACC-3 study endpoints being RFS and OS. RFS was defined as time in months from randomization until occurrence of local, regional or distant relapse, a second primary colon cancer or death. OS was defined as time in months from randomization until death. For ABCG2, the median from the whole expression data set (Stage II and III) was chosen to dichotomize the patients into ABCG2 high and ABCG2 low. For TOP1 mRNA, we used, based on our previous study [7], the third quartile on the whole expression data of stage II and III patients to group the patients into TOP1 high and low. ABCG2 and TOP1 were then combined into a binary variable: ABCG2 high (above median) and TOP1 low (below the 75-percentile) formed the "resistant patients", while all other combinations of ABCG2 and TOP1 formed the "sensitive patients ". The patients were further stratified according to the treatment (5FUL or FOLFIRI).

The Kaplan–Meier method was used to estimate RFS and OS rates, and univariate comparisons were made using the log rank test. The effect size of ABCG2/TOP1 status and treatment arm were estimated in univariate and multivariable analysis using the Cox proportional hazards model. Adjustment variables for multivariable analysis were selected based on LASSO penalized proportional hazards regression [9]. Microsatellite instability (MSI) data were available from a previous study [10] and were tested alongside the clinical and pathological baseline variables: N stage, tumor localization, tumor grade, sex, and age. Formal tests for statistical interaction between dichotomized ABCG2/TOP1 status ("resistant patients" vs "sensitive patients") and treatment were performed in separate Cox models, including main effects and an interaction term. All results were summarized in terms of hazard ratios (HR), estimated 95% confidence intervals (CI), and p-values from the Wald-test.

Pearson correlation coefficients (r) were calculated to test for statistical dependence between the ABCG2/TOP1 variables.

All p-values were two-sided and the significance level was set at 0.05. All analyses were performed in R software for statistical computing version 3.4.0 [17].

4.4. Subgroup Analyses

Since mechanisms of drug resistance effective in MSI tumors might be different from those in MSS tumors [18], we also divided the Stage III patients into MSS and MSI genotypes, (n = 470 tumors being MSS and MSI-L and n = 51 tumors being MSI-H (59 missing values)). Kaplan Meier survival statistics were used to estimate RFS and OS rates in each group according to ABCG2/TOP1 status dichotomized as described above.

The REMARK guidelines [19] were followed wherever applicable.

5. Conclusions

In conclusion, we show that ABCG2/TOP1 status as a combined test results is a potential biomarker, which provides significant predictive information on benefit of adjuvant irinotecan treatment of Stage III CC patients. However, our data could only show a trend for a better patient outcome with FOLFIRI treatment of "sensitive patients" as compared to the 5FUL treated patients. The predictive value of our biomarker test needs to be confirmed in an independent validation cohort. Our results also raise the question whether FOLFIRI biomarker positive patients will benefit from FOLFIRI only or whether they are those benefiting from adjuvant treatment with FOLFOX as well. An adjuvant study enrolling Stage III CC patients with a FOLFIRI "sensitive" gene profile and randomizing these patients to treatment with FOLFOX or FOLFIRI will answer this question.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/12/4/977/s1, Figure S1: Scatter plot of log2-expression levels of ABCG2 and TOP1 genes. Figure S2: RFS stratified by biomarker in whole stage III. Figure S3: RFS stage III stratified by biomarker and treatment. Figure S4: RFS in FOLFIRI by biomarker vs whole 5FUL arm. Figure S5: OS by biomarker in whole Stage III. Figure S6: OS by biomarker in FOLFIRI arm vs 5FUL arm. Figure S7: RFS stage III/MSS stratified by biomarker and treatment. Figure S8: OS for stage III/MSS stratified by biomarker and treatment. Multivariable Cox regression analyses.

Author Contributions: Conceptualization, J.S., E.B., V.P. and N.B.; methodology, J.S., E.B., V.P. and N.B.; validation, V.P.; formal analysis, J.S., E.B., V.P. and N.B.; investigation, J.S., E.B., E.v.C., F.B., V.P. and N.B.; resources, E.B, V.P. and N.B.; data curation, E.B., E.v.C., F.B. and V.P.; writing—original draft preparation, J.S., E.B., E.v.C., F.B., V.P. and N.B.; writing—review and editing, J.S., E.B., V.P. and N.B.; visualization, J.S., E.B., V.P. and N.B.; supervision, J.S., E.B., V.P. and N.B.; project administration, E.B., V.P. and N.B.; funding acquisition, E.v.C., F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Savvaerksejer Jeppe Juhl og Hustru Ovita Juhls Mindelegat, grant file number 102-5212/18-3000, the Agentura Pro Zdravotnický Výzkum České Republiky, grant file number NV19-03-00298, IMK Almene, grant number 102-0489/13-3000. The APC was funded by Scandion Oncology.

Acknowledgments: E.B. and V.P. acknowledge the support of RECETOX research infrastructure through project LM2018121 funded by the Czech Ministry of Education, Youth and Sports, and the support of CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 (857560) and the Czech Ministry of Education, Youth and Sports (02.1.01/0.0/0.0/18_046/0015975)".

Conflicts of Interest: The authors declare no conflict of interest and the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. André, T.; Boni, C.; Navarro, M.; Tabernero, J.; Hickish, T.; Topham, C.; Bonetti, A.; Clingan, P.; Bridgewater, J.; Rivera, F.; et al. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J. Clin. Oncol.* **2009**, *27*, 3109–3116. [CrossRef] [PubMed]
2. Tournigand, C.; André, T.; Achille, E.; Lledo, G.; Flesh, M.; Mery-Mignard, D.; Quinaux, E.; Couteau, C.; Buyse, M.; Ganem, G.; et al. FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: A randomized GERCOR study. *J. Clin. Oncol.* **2004**, *22*, 229–237. [CrossRef] [PubMed]
3. Van Cutsem, E.; Labianca, R.; Bodoky, G.; Barone, C.; Aranda, E.; Nordlinger, B.; Topham, C.; Tabernero, J.; André, T.; Sobrero, A.F.; et al. Randomized phase III trial comparing biweekly infusional

fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J. Clin. Oncol.* **2009**, *27*, 3117–3125. [CrossRef] [PubMed]

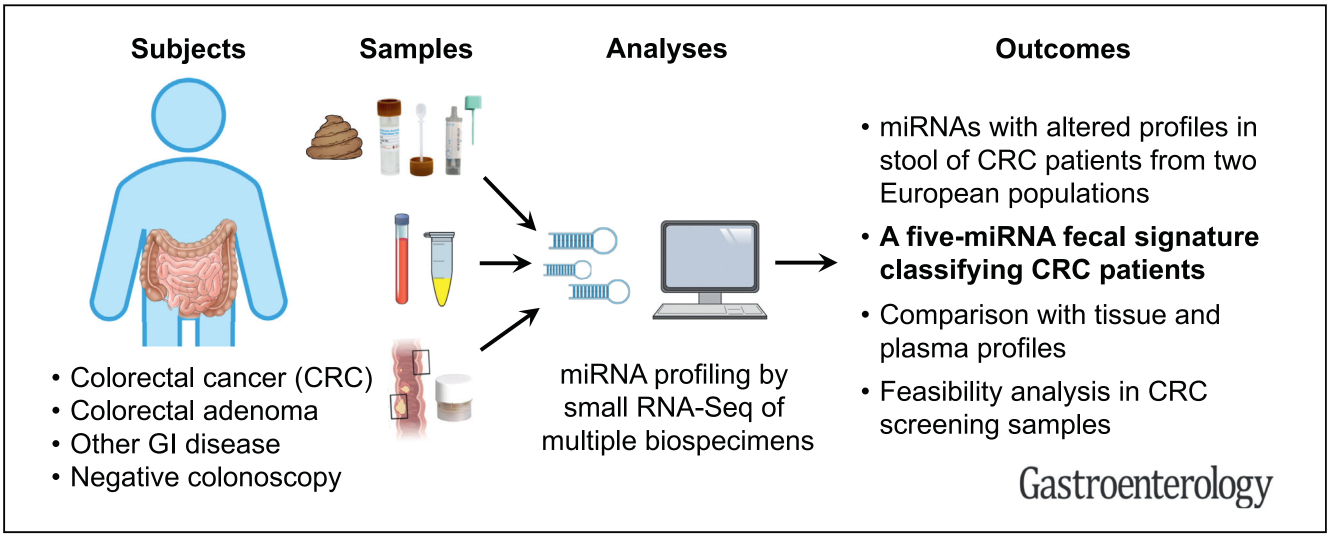
4. Saltz, L.B.; Niedzwiecki, D.; Hollis, D.; Goldberg, R.M.; Hantel, A.; Thomas, J.P.; Fields, A.L.; Mayer, R.J. Irinotecan fluorouracil plus leucovorin is not superior to fluorouracil plus leucovorin alone as adjuvant treatment for stage III colon cancer: Results of CALGB 89803. *J. Clin. Oncol.* **2007**, *25*, 3456–3461. [CrossRef] [PubMed]
5. Jensen, N.F.; Stenvang, J.; Beck, M.K.; Hanáková, B.; Belling, K.C.; Do, K.N.; Viuff, B.; Nygård, S.B.; Gupta, R.; Rasmussen, M.H.; et al. Establishment and characterization of models of chemotherapy resistance in colorectal cancer: Towards a predictive signature of chemoresistance. *Mol. Oncol.* **2015**, *9*, 1169–1185. [CrossRef] [PubMed]
6. Jandu, H.; Aluzaite, K.; Fogh, L.; Thrane, S.W.; Noer, J.B.; Proszek, J.; Do, K.N.; Hansen, S.N.; Damsgaard, B.; Nielsen, S.L.; et al. Molecular characterization of irinotecan (SN-38) resistant human breast cancer cell lines. *BMC Cancer* **2016**, *16*, 34. [CrossRef] [PubMed]
7. Nygård, S.B.; Vainer, B.; Nielsen, S.L.; Bosman, F.; Tejpar, S.; Roth, A.; Delorenzi, M.; Brünner, N.; Budinska, E. DNA Topoisomerase I Gene Copy Number and mRNA Expression Assessed as Predictive Biomarkers for Adjuvant Irinotecan in Stage II/III Colon Cancer. *Clin. Cancer Res.* **2016**, *22*, 1621–1631. [CrossRef] [PubMed]
8. Popovici, V.; Budinska, E.; Tejpar, S.; Weinrich, S.; Estrella, H.; Hodgson, G.; Van Cutsem, E.; Xie, T.; Bosman, F.T.; Roth, A.D.; et al. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *J. Clin. Oncol.* **2012**, *30*, 1288–1295. [CrossRef] [PubMed]
9. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **1997**, *16*, 385–395. [CrossRef]
10. Klingbiel, D.; Saridaki, Z.; Roth, A.D.; Bosman, F.T.; Delorenzi, M.; Tejpar, S. Prognosis of stage II and III colon cancer treated with adjuvant 5-fluorouracil or FOLFIRI in relation to microsatellite status: Results of the PETACC-3 trial. *Ann. Oncol.* **2015**, *26*, 126–132. [CrossRef] [PubMed]
11. Nielsen, D.L.; Palshof, J.A.; Brunner, N.; Stenvang, J.; Viuff, B.M. Implications of ABCG2 Expression on Irinotecan Treatment of Colorectal Cancer Patients: A Review. *Int. J. Mol. Sci.* **2017**, *18*, 1926. [CrossRef] [PubMed]
12. Ambjørner, S.E.; Wiese, M.; Köhler, S.C.; Svindt, J.; Lund, X.L.; Gajhede, M.; Saaby, L.; Brodin, B.; Rump, S.; Weigt, H.; et al. The Pyrazolo[3,4-d]pyrimidine Derivative, SCO-201, Reverses Multidrug Resistance Mediated by ABCG2/BCRP. *Cells* **2020**, *9*, 613. [CrossRef] [PubMed]
13. Stenvang, J.; Lima, T.; Nielsen, S.L.; Drejer, J.; Brunner, N.; Christophersen, P. The volume regulated anion channel inhibitor NS3728 to enhance the cytotoxic effects of SN-38 in human colorectal cancer cells grown in vitro. *J. Clin. Oncol.* **2016**, *34*, e23170. [CrossRef]
14. Cederbye, C.N.; Palshof, J.A.; Hansen, T.P.; Duun-Henriksen, A.K.; Linnemann, D.; Stenvang, J.; Nielsen, D.L.; Brünner, N.; Viuff, B.M. Antibody validation and scoring guidelines for ABCG2 immunohistochemical staining in formalin-fixed paraffin-embedded colon cancer tissue. *Sci. Rep.* **2016**, *6*, 26997. [CrossRef] [PubMed]
15. Simon, R.M.; Paik, S.; Hayes, D.F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.* **2009**, *101*, 1446–1452. [CrossRef] [PubMed]
16. Budinska, E.; Popovici, V.; Tejpar, S.; D'Ario, G.; Lapique, N.; Sikora, K.O.; Di Narzo, A.F.; Yan, P.; Hodgson, J.G.; Weinrich, S.; et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **2013**, *231*, 63–76. [CrossRef] [PubMed]
17. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
18. Sønderstrup, I.M.; Nygård, S.B.; Poulsen, T.S.; Linnemann, D.; Stenvang, J.; Nielsen, H.J.; Bartek, J.; Brünner, N.; Nørgaard, P.; Riis, L. Topoisomerase-1 and -2A gene copy numbers are elevated in mismatch repair-proficient colorectal cancers. *Mol. Oncol.* **2015**, *9*, 1207–1217. [CrossRef] [PubMed]
19. McShane, L.M.; Altman, D.G.; Sauerbrei, W.; Taube, S.E.; Gion, M.; Clark, G.M. Reporting recommendations for tumor marker prognostic studies (REMARK). *J. Natl. Cancer Inst.* **2005**, *97*, 1180–1184. [CrossRef] [PubMed]

[***19***] Pardini B, Ferrero G, Tarallo S, Gallo G, Francavilla A, Licheri N, Trompetto M, Clerico G, Senore C, Peyre S, Vymetalkova V, Vodickova L, Liska V, Vycital O, Levy M, Macinga P, Hucl T, **Budinska E,** Vodicka P, Cordero F, Naccarati A. A Fecal MicroRNA Signature by Small RNA Sequencing Accurately Distinguishes Colorectal Cancers: Results From a Multicenter Study. Gastroenterology. 2023 Sep;165(3):582-599.e8. doi: 10.1053/j.gastro.2023.05.037. Epub 2023 May 30. PMID: 37263306.

A Fecal MicroRNA Signature by Small RNA Sequencing Accurately Distinguishes Colorectal Cancers: Results From a Multicenter Study

Barbara Pardini,[1,2,*] **Giulio Ferrero**,[3,4,*] **Sonia Tarallo**,[1,2,*] Gaetano Gallo,[5,6] Antonio Francavilla,[1] Nicola Licheri,[4] Mario Trompetto,[6] Giuseppe Clerico,[6] Carlo Senore,[7] Sergio Peyre,[8] Veronika Vymetalkova,[9,10,11] Ludmila Vodickova,[9,10,11] Vaclav Liska,[11,12] Ondrej Vycital,[11,12] Miroslav Levy,[13] Peter Macinga,[14] Tomas Hucl,[14] Eva Budinska,[15] Pavel Vodicka,[9,10,11] **Francesca Cordero**,[4,§] and **Alessio Naccarati**[1,2,§]

[1]Italian Institute for Genomic Medicine, Turin, Italy; [2]Candiolo Cancer Institute, FPO-IRCCS, Turin, Italy; [3]Department of Clinical and Biological Sciences, University of Turin, Turin, Italy; [4]Department of Computer Science, University of Turin, Turin, Italy; [5]Department of Surgery, Sapienza University of Rome, Rome, Italy; [6]Department of Colorectal Surgery, Clinica S. Rita, Vercelli, Italy; [7]Epidemiology and Screening Unit-CPO, University Hospital Città della Salute e della Scienza, Turin, Italy; [8]LILT (Lega Italiana Lotta contro i Tumori), associazione provinciale di Biella, Biella, Italy; [9]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic; [10]Institute of Biology and Medical Genetics, 1st Medical Faculty, Charles University, Prague, Czech Republic; [11]Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic; [12]Department of Surgery, University Hospital and Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic; [13]Department of Surgery, First Faculty of Medicine, Charles University and Thomayer Hospital, Prague, Czech Republic; [14]Department of Gastroenterology and Hepatology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic; and [15]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic


Subjects
Samples
Analyses
Outcomes
• Colorectal cancer (CRC)
• Colorectal adenoma
• Other GI disease
• Negative colonoscopy
miRNA profiling by small RNA-Seq of multiple biospecimens
• miRNAs with altered profiles in stool of CRC patients from two European populations
• A five-miRNA fecal signature classifying CRC patients
• Comparison with tissue and plasma profiles
• Feasibility analysis in CRC screening samples
Gastroenterology

BACKGROUND & AIMS: Fecal tests currently used for colorectal cancer (CRC) screening show limited accuracy in detecting early tumors or precancerous lesions. In this respect, we comprehensively evaluated stool microRNA (miRNA) profiles as biomarkers for noninvasive CRC diagnosis. **METHODS:** A total of 1273 small RNA sequencing experiments were performed in multiple biospecimens. In a cross-sectional study, miRNA profiles were investigated in fecal samples from an Italian and a Czech cohort (155 CRCs, 87 adenomas, 96 other intestinal diseases, 141 colonoscopy-negative controls). A predictive miRNA signature for cancer detection was defined by a machine learning strategy and tested in additional fecal samples from 141 CRC patients and 80 healthy volunteers. miRNA profiles were compared with those of 132 tumors/adenomas paired with adjacent mucosa, 210 plasma extracellular vesicle samples, and 185 fecal immunochemical test leftover samples. **RESULTS:** Twenty-five miRNAs showed altered levels in the stool of CRC patients in both cohorts (adjusted $P < .05$). A 5-miRNA signature, including miR-149-3p, miR-607-5p, miR-1246, miR-4488, and miR-6777-5p, distinguished patients from control individuals (area under the curve [AUC], 0.86; 95% confidence interval [CI], 0.79–0.94) and was validated in an independent cohort (AUC, 0.96; 95% CI, 0.92–1.00). The signature classified control individuals from patients with low-/high-stage tumors and advanced adenomas (AUC, 0.82; 95% CI, 0.71–0.97). Tissue miRNA profiles mirrored those of stool samples, and fecal profiles of different gastrointestinal diseases highlighted miRNAs specifically dysregulated in CRC. miRNA profiles in fecal immunochemical test leftover samples showed good correlation with those of stool collected in preservative buffer, and their alterations could be detected in adenoma or CRC patients. **CONCLUSIONS:** Our comprehensive fecal miRNome analysis identified a signature accurately discriminating cancer aimed at improving noninvasive diagnosis and screening strategies.

Keywords: Stool MicroRNAs; Noninvasive Diagnosis; Small RNA Sequencing; Colorectal Cancer; Precancerous Lesions; Machine Learning.

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Current screening programs for the noninvasive detection of colorectal cancer (CRC) are based on fecal tests with limited accuracy for early malignancies or precancerous lesions. Evaluating microRNA (miRNA) profiles in stool could improve the screening strategy.

NEW FINDINGS

Investigating the whole miRNome in stool and with ad hoc explainable machine learning, we identified in 2 independent cohorts 5 miRNAs that could accurately classify CRC patients from control individuals. The signature was validated in a third cohort and assayed in fecal immunochemical test leftover samples from the screening.

LIMITATIONS

Despite the large number of samples overall collected and sequenced, the disease subtypes investigated were still not exhaustive of the heterogeneity in CRC and adenomas. Although we showed the feasibility of the molecular analysis, the investigation on screening samples still represents a pilot approach.

CLINICAL RESEARCH RELEVANCE

The investigation of the whole miRNome in all of the cohorts led to a comprehensive overview of the fecal miRNA profiles, providing the possibility to accurately single out those signals that may enhance the accuracy of the screening. The identified miRNA signature accurately discriminates different stages of CRC development, and it constitutes a coadjuvant to current screening programs for a noninvasive, accurate diagnosis.

BASIC RESEARCH RELEVANCE

New and previously reported miRNAs altered in CRC are detectable in stool and may highlight a novel role of these molecules released in the gut in physiologic and pathologic conditions.

In the last 30 years, we have witnessed a dramatic increase in understanding the epidemiology, etiology, molecular biology, and various clinical aspects of colorectal cancer (CRC).[1] However, approximately 1.8 million new cases are annually diagnosed worldwide, posing CRC as the third most common incident cancer. Moreover, although early-stage tumors can be efficiently treated, CRC is still the second-leading cause of cancer-related death, with 900,000 deaths in 2018.[2,3] Hence, the early detection of preclinical cancers or precursor lesions is a desirable objective, because it may strongly increase the chances for successful treatment and cure.

Most European countries have implemented CRC screening programs based on noninvasive stool tests for detecting fecal occult blood, mainly the fecal immunochemical test (FIT).[4,5] FIT selects individuals showing a higher prevalence of CRC and advanced benign neoplasia but has limited sensitivity to recognize advanced colorectal adenomas (AAs).[6] Colonoscopy is also used in an opportunistic screening setting and detects both cancer and premalignant lesions but is bothersome and invasive, as well as costly for the health system.[7] Despite the fact that FIT-based screening programs are undeniably efficient in detecting premalignant growths and providing an earlier diagnosis, successfully reducing CRC burden, only approximately 5% of individuals who receive a colonoscopy based on FIT results will end up with a significant lesion (CRC or AA). Stool tests show a relatively low specificity, resulting in a high number of false positives and a considerable number of unnecessary colonoscopies.[8] Complementing traditional screening stool tests with other noninvasively detectable fecal molecular biomarkers could improve the triage of individual for colonoscopy, reducing the costs for the health systems in terms of the number of examinations and decreasing the risks and discomfort for patients.[9,10]

Identifying reliable biomarkers is not trivial, given the ensemble of hidden interactions between molecules and patient-specific clinical/anamnestic characteristics. However, machine learning (ML) algorithms have been defined to reveal significant features able to accurately discriminate groups of individuals. In particular, explainable ML approaches allow the identification of novel molecular biomarker signatures to improve early CRC diagnosis, as recently demonstrated for fecal microbial species[11] and urinary proteins.[12]

The analysis of small noncoding RNAs in fecal samples has attracted interest with an excellent biological and analytic rationale for its application in large-scale clinical investigations.[13] Tumor-secreted small noncoding RNAs are directly and continuously released into the intestinal lumen,

* Authors share co-first authorship; § Authors share co-senior authorship.

Abbreviations used in this paper: AA, advanced colorectal adenoma; AUC, area under the curve; BMI, body mass index; CRC, colorectal cancer; CZ, Czech; DEmiRNA, differentially expressed microRNA; EV, extracellular vesicle; FIT, fecal immunochemical test; GI, gastrointestinal; IBD, inflammatory bowel disease; IT, Italian; miRNA, microRNA; ML, machine learning; nAA, nonadvanced adenoma; qPCR, quantitative polymerase chain reaction; RNA-seq, RNA sequencing; RT, reverse transcription.

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2023.05.037

and their profiles may be altered in concomitance with the presence of CRC and precancerous lesions. Moreover, small noncoding RNAs, such as microRNAs (miRNAs), are remarkably stable, enabling their accurate detection in stool without the need for special stabilization or logistic requirements.[14] miRNAs are suitable biomarkers in surrogate tissues and biofluids because their levels are altered in specific pathologic states,[15] in the presence of precursor lesions,[16] and in CRC development.[17–19] In addition, specific fecal miRNA alterations have been associated with the gut microbiome composition[20] and proposed as noninvasive CRC biomarkers.[21]

So far, comprehensive miRNA profiling by small RNA sequencing (small RNA-seq) has been mainly performed on tumor tissues or plasma.[21,22] In contrast, studies on fecal samples investigated few miRNAs in relation to CRC, typically in small cohorts and without taking into account their demographic characteristics.[23] In this respect, studies on the whole fecal miRNome showed that different lifestyles and dietary habits might critically affect specific miRNA levels.[24,25] In addition, limited evidence is available on stool miRNA profiles in relation to patient clinicopathologic characteristics, such as specific CRC stages, precancerous lesions or other gastrointestinal (GI) diseases, except for the reported pleiotropic dysregulation of miR-21-5p in several diseases.[26] Therefore, an miRNA signature for CRC detection derived from a comprehensive fecal miRNome analysis across multiple populations is currently lacking.

This multicenter study aimed to explore, by deep sequencing, the miRNA profiles in stool samples that best characterize CRC patients from control individuals and distinguish colorectal adenomas or other GI diseases. The analyses were performed in different independent cohorts adopting the same protocol for participant recruitment, sample collection, and small RNA-seq experiments/analyses. An integrated explainable ML strategy identified a fecal miRNA signature distinguishing CRC patients from control individuals, and the results were validated in an additional cohort. Finally, altered miRNAs in stool were also investigated in FIT-positive leftover samples collected within a population-based CRC screening program.

Methods

Stool Study Cohorts

Italian cohort. Stool specimens as well as clinical and demographic data were collected from 317 individuals recruited in a hospital-based study in Vercelli, Italy (Table 1). Based on the results of complete colonoscopy examination, participants were classified into (1) 89 sporadic CRC patients, (2) 74 polyp patients (6 hyperplastic polyps, 20 nonadvanced adenomas [nAAs] and 48 AAs; serrated lesions were excluded because there were too few), (3) 49 individuals with a GI disease (6 Crohn's disease, 9 ulcerative colitis, 14 diverticulitis, 7 diverticulosis, 13 hemorrhoidal disease), and (4) 105 colonoscopy-negative control individuals. AAs were defined based on the presence of high-grade dysplasia, villous component, or lesion size of >1 cm as defined by Zarchy and Ershoff.[27] Of this cohort, 93 stool samples (from 29 CRC patients, 27 polyps, 13 patients with a GI disease, and 24 colonoscopy-negative control individuals) have been used and described previously in other studies.[11,28,29]

Czech cohort. Stool specimens as well as clinical and demographic data were collected from 162 Czech individuals recruited in 2 hospitals in Prague and 1 in Plzen, Czech Republic (Table 1). Based on colonoscopy results, participants were divided in (1) 66 CRC patients, (2) 28 polyp patients (9 hyperplastic polyps, 13 nAAs, 6 AAs; no serrated lesions were collected), (3) 32 patients with other GI diseases (3 Crohn's disease, 11 ulcerative colitis, 17 diverticulosis, 1 unclassified inflammatory bowel disease [IBD]); and (4) 36 colonoscopy-negative individuals.

Validation cohort. Stool specimens from 141 CRC patients recruited in a hospital in Brno, Czech Republic, and 80 stool samples of healthy volunteers contributing to science were included. These participants were previously described in other studies: the CRC population is described by Zwinsova et al[30] but here is sequenced for the first time for small RNA-seq; healthy volunteers are a part of the cohorts described and sequenced for small noncoding RNAs by Tarallo et al[24] and Francavilla et al.[31]

Fecal immunochemical test cohort. FIT buffer leftover samples from 185 individuals with a positive test result were collected within the CRC screening for the Piedmont Region (Italy). Based on colonoscopy results, participants were classified as control individuals (n = 53), AA (n = 80), nAA (n = 30), or CRC (n = 22). Among them, 57 individuals also provided stool samples before undergoing colonoscopy.

More details on the cohorts included in the study are given in the Supplementary Materials. The local ethics committees of Azienda Ospedaliera SS. Antonio e Biagio e C. Arrigo of Alessandria (Italy, protocol no. Colorectal miRNA CEC2014), AOU Città della salute e della Scienza di Torino (Italy), the Institute of Experimental Medicine of Prague (Czech Republic), Masaryk Memorial Cancer Institute (protocol no. 2018/865/MOU), and Masaryk University of Brno (Czech Republic, protocol no. EKV-2019-044) approved the study. All patients gave written informed consent following the Declaration of Helsinki before participating in the study.

Other Analyzed Biospecimens

For 132 patients having surgery at the Vercelli hospital, primary tissues (102 CRC and 30 adenomas) paired with adjacent colonic mucosa were collected.

Blood samples were collected from 210 out of 317 Italian (IT) cohort participants, stratified into patients with CRC (n = 52), AAs (n = 19), nAAs (n = 15), hyperplastic polyps (n = 6), and other GI diseases (n = 39), and control individuals (n = 79).

Sample Collection

Naturally evacuated fecal samples were obtained from participants previously instructed to self-collect the specimen at home. Samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp). Stool aliquots (200 μL) were stored at –80°C until RNA extraction.[20] For the validation cohort of CRC patients from Brno, stool samples were collected from untreated patients before the scheduled surgery with DNA-free swabs (Deltalab). Patients performed the collection at home and returned the samples to the hospital, where they were immediately frozen at –80°C until further processing.

Table 1. Study Population Characteristics

	IT cohort (n = 317)					CZ cohort (n = 162)				
Covariate	Controls (n = 105)	Other GI disease (n = 49)	Polyps (n = 74)	CRC (n = 89)	*P* value	Controls (n = 36)	Other GI disease (n = 32)	Polyps (n = 28)	CRC (n = 66)	*P* value
Age, *y*										
Average ± SD	59.6 ± 10.7	56.7 ± 13.6	66.2 ± 9.1	70.6 ± 9.7	7.34E–13	57.8 ± 10.5	58.7 ± 9.4	63.1 ± 8.4	68.0 ± 11.2	8.34E–06
Range	39–84	30–82	42–93	50–88		40–76	41–75	48–82	40–88	
Sex, n										
Male	52	23	41	52	4.83E–01	14	16	14	46	1.74E–02
Female	53	26	33	37		22	16	14	20	
BMI, kg/m^2										
Average	25.3 ± 4.5	25.0 ± 3.4	25.0 ± 3.7	25.8 ± 5.1	9.02E–01	28.2 ± 6.1	28.8 ± 7.0	29.0 ± 3.5	27.1 ± 5.4	1.61E–01
Range	15.4–40.0	19.5–33.7	19.5–36.0	16.0–44.1		21.0–43.9	22.0–60.9	22.6–34.7	16.9–47.6	
Smoking status, n										
Nonsmoker	31	17	18	35	2.16E–01	25	24	13	32	2.53E–02
Ex-smoker	16	6	20	15		3	0	8	12	
Smoker	38	12	22	31		8	8	6	18	
NA	20	14	14	7		0	0	1	4	
Localization, n[a]										
Proximal			19	37				16	16	
Distal			11	20				11	15	
Rectum			18	28				6	34	
NA			32	6				0	1	
Polyp type, n										
Tubular adenoma			18					19		
Tubulovillous adenoma			12					0		
Tubular sessile			5					0		
Hyperplastic polyp			6					9		
NA			31					0		
Adenoma type, n										
AA			48					6		
nAA			20					13		
pT (combined), n										
T1–T2				27					20	
T3–T4				54					43	
Tis				0					1	
NA				7					2	

Table 1. Continued

	IT cohort (n = 317)					CZ cohort (n = 162)				
Covariate	Controls (n = 105)	Other GI disease (n = 49)	Polyps (n = 74)	CRC (n = 89)	*P* value	Controls (n = 36)	Other GI disease (n = 32)	Polyps (n = 28)	CRC (n = 66)	*P* value
AJCC staging, n										
I				18					16	
II				24					16	
III				29					15	
IV				5					14	
NA				13					5	
Grade, n										
G1–G2				39					44	
G3				38					18	
NA				12					4	
Metastasis (lymph node or distal), n										
No				49					52	
Yes				31					11	
NA				9					3	
Other GI diseases, n										
Crohn's disease		6					3			
Ulcerative rectocolitis		9					11			
Diverticulosis		7					17			
Diverticulitis		14					0			
Hemorrhoidal disease		13								
NA		0					1			

AJCC, American Joint Committee on Cancer; NA, not available; pT, post-operatory tumor size; SD, standard deviation.
[a]Totals may be different from the total number of individuals in each category because of the presence of multiple lesions.

For the FIT cohort, leftovers from FIT tubes (~1.2 mL) used for automated tests (OC-sensor, Eiken Chemical Co) for hemoglobin quantification were stored at –80°C until use.

Plasma samples were obtained from 8 mL of blood centrifuged for 10 minutes at 1000 revolutions/minute, and aliquots were stored at –80°C until use. Plasma extracellular vesicles (EVs) were precipitated from 200 μL of plasma using ExoQuick (System Biosciences) according to Sabo et al.[32]

Paired tumor/adenoma tissue and adjacent nonmalignant mucosa (at least 20 cm distant) were obtained from CRC and adenoma patients during surgical resection and immediately immersed in RNAlater solution (Ambion). All samples were stored at –80°C until use.

Total RNA Extraction, Small RNA Sequencing Library Preparation, and Quantitative Real-Time Polymerase Chain Reaction

Total RNA from stool and FIT leftover samples was extracted using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as previously described.[20] Total RNA from plasma EVs was extracted as described in Sabo et al.[32] For tissue samples, total RNA was extracted using QIAzol (Qiagen) according to the manufacturer's instructions.

Small RNAs were converted into barcoded complementary DNA libraries for Illumina single-end sequencing (75 cycles on HiSeq4000 or NextSeq500, Illumina Inc) as previously described.[24]

Candidate miRNA biomarkers were replicated in stool samples using the miRCURY LNA miRNA PCR Assays (Qiagen). Reverse transcription (RT) was performed using the miRCURY LNA RT kit (Qiagen) according to the manufacturer's instructions. All reactions were run on an ABI Prism 7900 Sequence Detection System (Applied Biosystems). Analyses were performed as described by Moisoiu et al.[33] More details are provided in the Supplementary Materials.

Computational and Statistical Analyses

Small RNA-seq analyses were performed as described by Tarallo et al,[20] considering a curated miRNA reference based on miRBase v22 and including a characterization of novel miRNAs (Supplementary Table 1*A*). Differential expression analyses were performed using DESeq2 v1.22.2.[34] Functional enrichment analysis was performed with RBiomirGS v0.2.12,[35] considering the validated miRNA-target interactions. A generalized linear model was defined by considering the miRNA levels as the dependent variable and participant age, sex, body mass index (BMI), smoking habit, and cohort as independent variables.

An ML strategy was implemented to identify the optimal fecal miRNA signature to accurately classify CRC patients from control individuals. The ML approach is composed of 3 phases: data preparation, feature selection, and classification. (More details are provided in the Supplementary Materials.) The signature was determined by considering an increasing number of miRNAs prioritized by filter and classifier-embedded methods applied to the training set (70% of the IT/Czech [CZ] cohorts). The optimal set of miRNAs providing the highest area under the curve (AUC) was selected and further tested by 100 stratified 10-fold cross-validations, first on the remaining 30% of the IT/CZ cohorts excluded from the training set and then on the validation cohort. The training and test sets were defined by a stratified selection to maintain the same proportion of participants characterized by specific covariates (ie, age, sex, cohort, disease status, and tumor staging).

Other statistical tests were performed using the Wilcoxon-Mann-Whitney and Kruskal-Wallis (continuous variables) or chi-square (categorical variables) methods. The Benjamini-Hochberg method was used for multiple-testing correction. Results were considered significant at $P < .05$.

Study Design

This study was designed to define and characterize a fecal miRNA signature that accurately distinguishes CRC patients from control individuals (Figure 1). The applied analysis strategy included the following phases.

Fecal miRNome profiling and biomarker discovery.

- Detection of stool miRNAs with altered levels in CRC: miRNA profiles from small RNA-seq and metadata were used for a differential expression analysis between CRC patients and control individuals of both the IT cohort and CZ cohort, independently. The overlapping differentially expressed miRNAs (DEmiRNAs) from both cohorts were the input of the next step.

- Feature selection and definition of an miRNA predictive signature: An ML strategy identified an miRNA signature composed of the minimal set of DEmiRNAs that better distinguished CRC patients from control individuals by a stratified cross-validation procedure.

- Validation of the miRNA predictive signature. The signature performance was estimated in the validation cohort by a stratified cross-validation procedure.

Fecal differentially expressed microRNA characterization in different sample types and diseases.

- Assessment of DEmiRNA profiles in different biospecimens and clinical situations: DEmiRNA levels were evaluated in (1) tumor/adenoma tissue and adjacent mucosa, (2) plasma EVs of CRC patients and control individual, and (3) fecal samples from patients with a GI disease or precancerous lesions to identify CRC-specific or commonly altered miRNAs. In particular, the miRNA signature from (1) was also tested in the discrimination of patients with precancerous lesions (AA or nAA), alone or in combination with CRC, from control individuals.

- Testing the DEmiRNA levels in samples from a CRC screening program: DEmiRNA profiles were explored in parallel in FIT buffer leftovers and in stool collected in tubes with RNA stabilizing solution. Subsequently, stool DEmiRNA levels were analyzed in the leftover samples of the FIT cohort by stratifying participants based on the colonoscopy results.

A detailed description of the methods is provided in the Supplementary Materials.

Results

Stool MicroRNA Profiles Are Altered in Colorectal Cancer Patients: Evidence From 2 European Populations

In agreement with previous studies,[20,24,31] an average of 479 (range, 86–1516) miRNAs were detected in each stool sample by small RNA-seq (further details in the Supplementary Materials and Supplementary Table 1*B* and *C*). The age- and sex-adjusted differential expression analysis between CRC patients and control individuals was performed independently on both the IT cohort and CZ cohort identifying, respectively, 250 and 29 DEmiRNAs (median expression, >20 reads; adjusted *P* < .05) (Figure 2*A* and Supplementary Table 2*A*).

Twenty-five stool DEmiRNAs were in common between both cohorts (Figure 2*B*, Table 2, and Supplementary Table 2*A*), all with a coherent expression trend (20 up-regulated and 5 down-regulated; rho = 0.75; *P* < .001) (Figure 2*B*). The alteration of these fecal miRNA levels in relation to CRC was further supported by a generalized linear model analysis adjusted for cohort, age, sex, BMI, and smoking habits: 22 out of the 25 DEmiRNAs remained significantly associated (*P* < .05) (Supplementary Table 2*B*). DEmiRNA profiles were further explored in relation to CRC patient clinical data (Figure 2*C* and *D*). The levels of 3 down-regulated miRNAs (miR-607-5p, miR-677-5p, and miR-922-5p) significantly decreased with increasing tumor size (Figure 2*D*). miR-922-5p also significantly decreased in patients with advanced disease stages or lymph node invasion (Figure 2*D* and Supplementary Table 2*C*). Conversely, increasing levels of 19 out of the 20 up-regulated miRNAs in CRC were observed along with tumor size, with miR-1246, miR-1290, miR-148-3p, and miR-194-5p significantly related to this parameter. The levels of 11 CRC–up-regulated miRNAs significantly increased in patients with lymph node invasion. In addition, the levels of 11 miRNAs were significantly higher in samples from patients with rectal compared to colon cancers (Figure 2*D*).

Functional analysis of DEmiRNA target genes showed their involvement in cancer-related processes, including cell cycle regulation and DNA repair, particularly for up-regulated miRNA targets (Supplementary Table 2*D* and *E*).

A Fecal MicroRNA Signature Distinguishes Colorectal Cancer Patients From Control Individuals

An explainable ML strategy was implemented to identify the minimal set of miRNAs as a signature for CRC detection (Supplementary Figure 1 and Supplementary Materials). The pipeline was applied on the 25 DEmiRNA profiles and considering 70% of the IT cohort and CZ cohort as the training set (Supplementary Table 3*A*). The best miRNA signature distinguishing CRC patients from control individuals included miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p, and miR-1246 (AUC, 0.87 ± 0.01) (Figure 2*E*). This set of 5 miRNAs represented the best combination of noncorrelated molecules with the highest discriminative power. Moreover, they showed a good performance in the classification of the 30% of participants excluded from the training set (AUC, 0.81 ± 0.01) (Figure 2*F*). The classification improved after the inclusion of sex and age in the model (AUC, 0.86 ± 0.01) (Table 3 and Supplementary Table 3*B*). The performance of the signature was again tested in the validation cohort, where it remained fairly similar, irrespective (AUC, 0.91 ± 0.01) or not (AUC, 0.96 ± 0.01) of age and sex (Figure 2*F*, Table 3, and Supplementary Table 3*B*).

By stratifying patients for CRC stage, the same 5-miRNA signature accurately distinguished patients with stages III–IV CRC (validation cohort: AUC, 0.96 ± 0.01 and 0.94 ± 0.01, respectively, including or not age and sex), or CRC stages I–II from control individuals (validation cohort: AUC, 0.95 ± 0.01 and 0.87 ± 0.01, respectively, including or not age and sex) (Table 3 and Supplementary Table 3*B*).

The panel of 5 miRNAs of the signature identified by sequencing was tested by RT quantitative polymerase chain reaction (qPCR) in RNA isolated from a subset of 96 stool samples equally distributed among IT and CZ cohort participants, with a balanced number of CRC patients and control individuals (Supplementary Figure 2*A*). The 5 miRNAs were detected in all samples, also using this second method. The normalized levels from RT-qPCR showed patterns comparable to those provided by sequencing, except for miR-4488 (Supplementary Figure 2*A*). In particular, miR-1246 and miR-149-3p levels were significantly increased in patient samples. The same method was used to test the 5 miRNA levels in RNA from 8 FIT leftover samples of participants with a positive FIT result at the CRC screening: all miRNAs were also detected in this biospecimen (data not shown).

For 4 signature miRNAs, a concordant expression pattern was observed between small RNA-seq and RT-qPCR normalized levels, particularly for miR-1246 (rho = 0.63, *P* < .001) and miR-149-3p (rho = 0.26, *P* < .05) (Supplementary Table 3*C* and Supplementary Figure 2*B*). Only the levels of miR-4488 were characterized by a negative correlation (rho = –0.48, *P* < .001) in CRC patients only.

Stool Differentially Expressed MicroRNA Profiles Mirror Those of Primary Colorectal Cancer and Adenoma Tissues

A paired differential expression analysis was performed between tumor tissues and matched adjacent mucosa collected from 102 CRC patients. Among the 25 stool DEmiRNAs, 14 were differentially expressed (adjusted *P* < .05) in this comparison (Figure 3*A* and Supplementary Table 4*A*), with 7 miRNAs (miR-21-5p, miR-1246, miR-1290, miR-148a-3p, miR-4488, miR-149-3p, miR-12114) up-regulated in tumor tissues coherently with their increase in CRC patient stool. Among them, 3 (miR-1246, miR-4488, miR-149-3p) were included in our miRNA signature. The 5 miRNAs significantly down-regulated in CRC patient stool (miR-607-5p, miR-6777-5p, included in the 5-miRNA signature; miR-6076; miR-922-5p; and miR-9899) were poorly expressed (normalized reads, <20) in both tumor and adjacent tissues (Supplementary Table 4*A*).

The differential analysis performed on 30 adenoma tissues matched with adjacent mucosa showed miR-21-5p, miR-1290, miR-148a-3p, and miR-200b-3p as significantly up-regulated in adenoma tissues (adjusted *P* < .001), whereas let-7i-5p and miR-4508 were down-regulated (Figure 3*A* and Supplementary Table 4*A*).

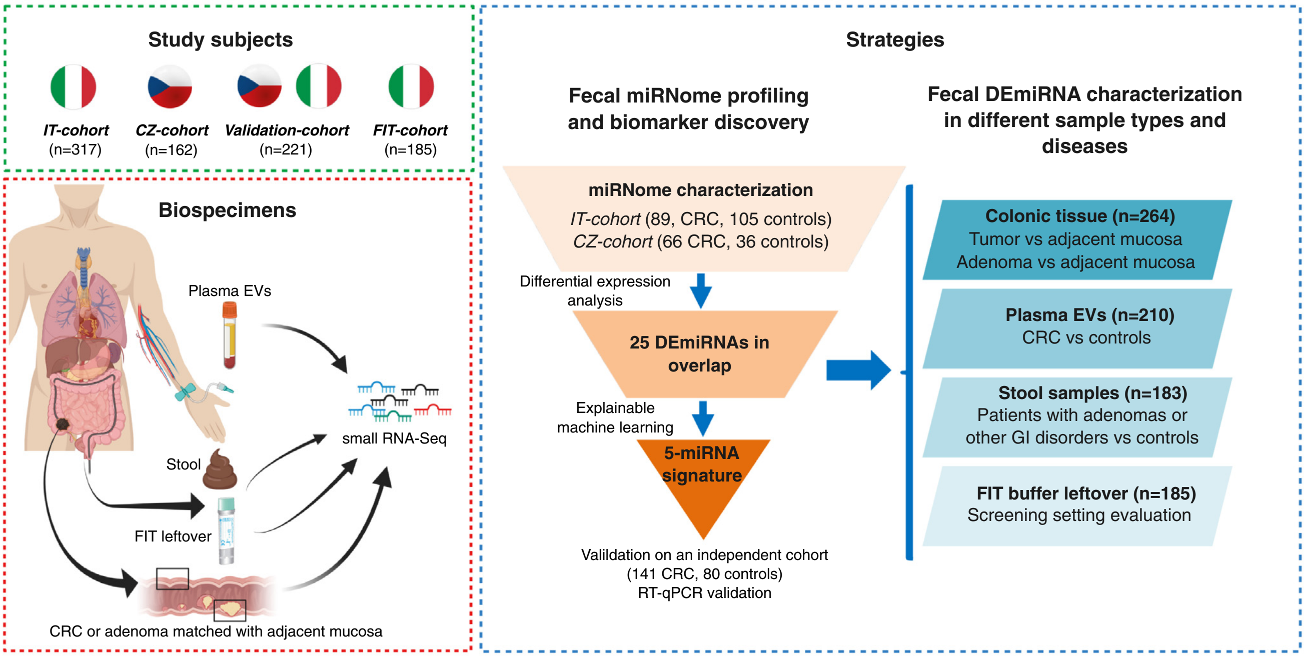

Study subjects
IT-cohort (n=317)
CZ-cohort (n=162)
Validation-cohort (n=221)
FIT-cohort (n=185)
Biospecimens
Plasma EVs
Stool
FIT leftover
small RNA-Seq
CRC or adenoma matched with adjacent mucosa
Strategies
Fecal miRNome profiling and biomarker discovery
miRNome characterization
IT-cohort (89, CRC, 105 controls)
CZ-cohort (66 CRC, 36 controls)
Differential expression analysis
25 DEmiRNAs in overlap
Explainable machine learning
5-miRNA signature
Valildation on an independent cohort (141 CRC, 80 controls) RT-qPCR validation
Fecal DEmiRNA characterization in different sample types and diseases
Colonic tissue (n=264)
Tumor vs adjacent mucosa
Adenoma vs adjacent mucosa
Plasma EVs (n=210)
CRC vs controls
Stool samples (n=183)
Patients with adenomas or other GI disorders vs controls
FIT buffer leftover (n=185)
Screening setting evaluation

Figure 1. Representation of the study design.

Few MicroRNA Levels Are Dysregulated in Circulating Extracellular Vesicles of Colorectal Cancer Patients

Small RNA-seq was performed on RNA isolated from plasma EVs collected from 210 participants in the IT cohort, detecting an average of 309 (range, 252–1213) miRNAs in these samples (Supplementary Table 4*B*). Among the 25 DEmiRNAs identified in stool samples of CRC patients, both miR-1246 and miR-4488 emerged as coherently significantly dysregulated in plasma EVs, although the latter was associated with low levels (normalized reads, <20) (Supplementary Table 4*B*). Another miRNA (miR-150-5p) was differentially expressed between CRC patients and control individuals (Supplementary Table 4*B*).

A Subset of Stool Differentially Expressed MicroRNAs Is Specifically Dysregulated in Colorectal Cancer Patients but Not in Those With Other GI Diseases

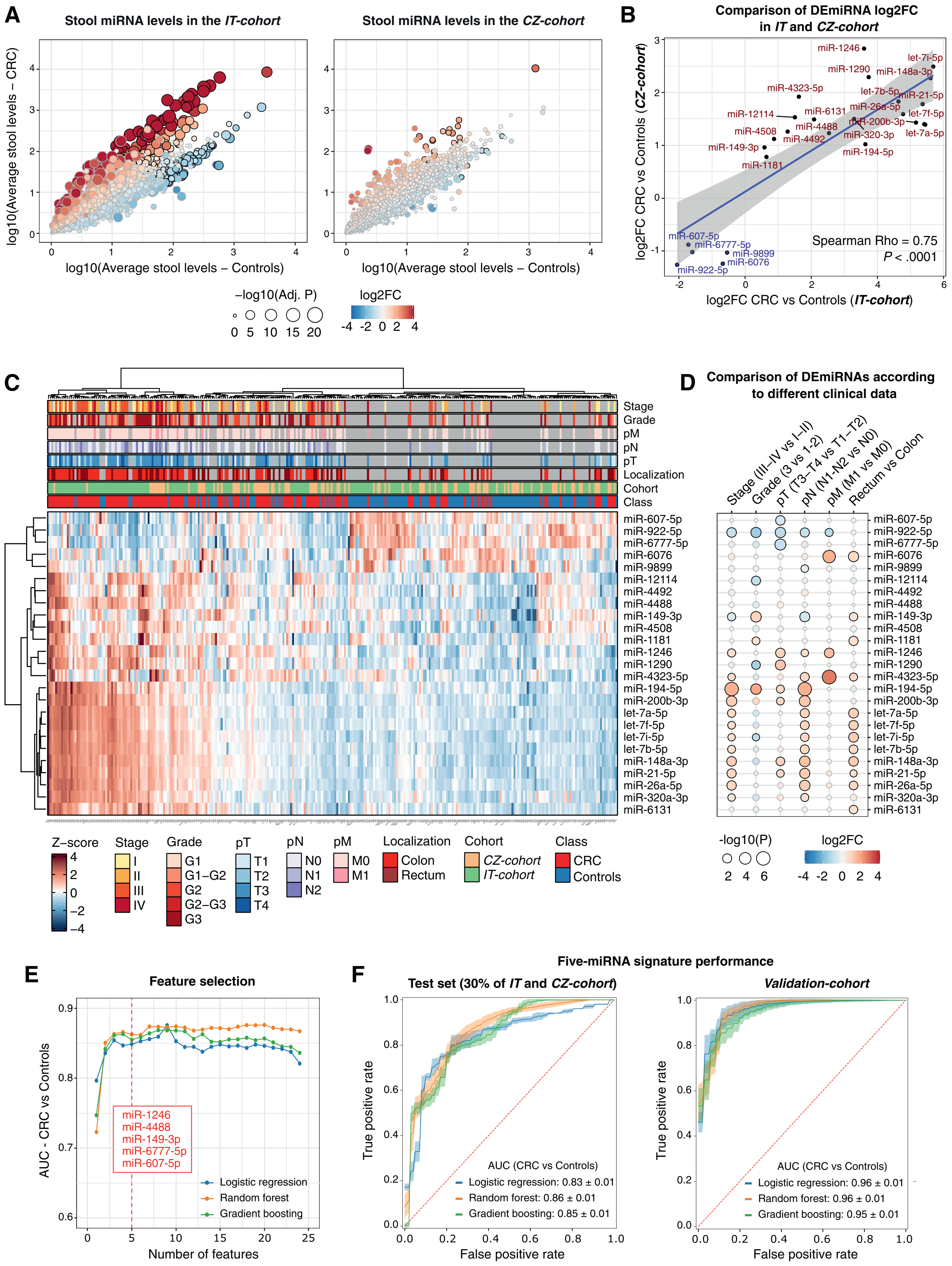
The CRC DEmiRNAs were further compared with those from patients with GI disorders and other precancerous lesions in both the IT and CZ cohorts. The age-, sex-, and cohort-adjusted differential expression analysis between each disease category and control individuals showed that the levels of 21 out of the 25 CRC DEmiRNAs were significantly altered in at least another GI disease (Figure 3*B*). Notably, in patients with ulcerative colitis, diverticulitis, nAA, or AA, 60% of the CRC DEmiRNAs were also dysregulated (Figure 3*B* and Supplementary Table 4*C*). The lowest number of dysregulated miRNAs was observed in patients with Crohn's disease (2 miRNAs) or diverticulosis (5 miRNAs), whereas no DEmiRNAs were found in patients with hyperplastic polyps.

Considering the 5 miRNAs constituting our predictive signature to distinguish CRC patients from control individuals, miR-6777-5p was not differentially expressed (compared to control individuals) in any other GI disease, miR-149-3p was significantly up-regulated only in patients with AA, and miR-607-5p was significantly down-regulated in patients with AA or ulcerative colitis compared to control individuals (Figure 3*B* and Supplementary Table 4*C*). Conversely, miR-4488 and miR-1246 stool levels significantly increased in patients with diverticulosis, ulcerative colitis, diverticulitis, or AA, with the latter miRNA also increased in Crohn's disease patients.

The identified signature was also used to classify AA and nAA patients from control individuals. Specifically, the miRNA signature was able to distinguish AA from control participants, both including (AUC, 0.82 ± 0.01) or not (AUC, 0.77 ± 0.02) age and sex in the analysis, as well as nAA (AUC, 0.80 ± 0.03 and 0.77 ± 0.02, respectively, including or not age and sex). Finally, patients with either CRC or AA were accurately distinguished from control individuals (including or not age and sex: AUC, 0.84 ± 0.01 and 0.81 ± 0.01, respectively) but not between them (CRC vs AA: AUC, 0.68 ± 0.02) (Table 3 and Supplementary Table 3*B*).

MicroRNAs Are Detectable in Fecal Immunochemical Test Leftover Samples by Small RNA Sequencing

The sequencing analysis was extended to 185 available leftover samples of the FIT cohort, still detecting an average of 618 miRNAs in each sample (Supplementary Table 1*B*). All of the 25 stool DEmiRNAs were detected in this type of sample. Considering the threshold adopted by our pipeline (ie, a minimum of 20 reads), 4 (miR-607-5p, miR-1246, let-


A
Stool miRNA levels in the IT-cohort
Stool miRNA levels in the CZ-cohort
log10(Average stool levels – CRC)
log10(Average stool levels – Controls)
−log10(Adj. P)
0 5 10 15 20
log2FC
-4 -2 0 2 4
B
Comparison of DEmiRNA log2FC in IT and CZ-cohort
log2FC CRC vs Controls (CZ-cohort)
log2FC CRC vs Controls (IT-cohort)
Spearman Rho = 0.75
P < .0001
miR-1246
let-7i-5p
miR-1290
miR-148a-3p
miR-4323-5p
let-7b-5p
miR-21-5p
miR-26a-5p
miR-12114
miR-6131
let-7f-5p
miR-200b-3p
miR-4508
miR-4488
miR-320-3p
let-7a-5p
miR-4492
miR-149-3p
miR-194-5p
miR-1181
miR-607-5p
miR-6777-5p
miR-9899
miR-6076
miR-922-5p
C
Stage
Grade
pM
pN
pT
Localization
Cohort
Class
miR-607-5p
miR-922-5p
miR-6777-5p
miR-6076
miR-9899
miR-12114
miR-4492
miR-4488
miR-149-3p
miR-4508
miR-1181
miR-1246
miR-1290
miR-4323-5p
miR-194-5p
miR-200b-3p
let-7a-5p
let-7f-5p
let-7i-5p
let-7b-5p
miR-148a-3p
miR-21-5p
miR-26a-5p
miR-320a-3p
miR-6131
Z–score
4 2 0 −2 −4
Stage
I II III IV
Grade
G1 G1−G2 G2 G2−G3 G3
pT
T1 T2 T3 T4
pN
N0 N1 N2
pM
M0 M1
Localization
Colon Rectum
Cohort
CZ-cohort IT-cohort
Class
CRC Controls
D
Comparison of DEmiRNAs according to different clinical data
Stage (III–IV vs I–II)
Grade (3 vs 1-2)
pT (T3–T4 vs T1–T2)
pN (N1-N2 vs N0)
pM (M1 vs M0)
Rectum vs Colon
-log10(P)
2 4 6
log2FC
-4 -2 0 2 4
E
Feature selection
AUC - CRC vs Controls
Number of features
miR-1246
miR-4488
miR-149-3p
miR-6777-5p
miR-607-5p
Logistic regression
Random forest
Gradient boosting
F
Five-miRNA signature performance
Test set (30% of IT and CZ-cohort)
Validation-cohort
True positive rate
False positive rate
AUC (CRC vs Controls)
Logistic regression: 0.83 ± 0.01
Random forest: 0.86 ± 0.01
Gradient boosting: 0.85 ± 0.01
AUC (CRC vs Controls)
Logistic regression: 0.96 ± 0.01
Random forest: 0.96 ± 0.01
Gradient boosting: 0.95 ± 0.01

7a-3p, miR-922) were detected in all samples, and 18 were detected in more than half (Figure 3*C* and Supplementary Table 4*D*). Three miRNAs included in our signature (miR-607-5p, miR-1246, miR-6777-5p) were detected in more than 95% of samples (Figure 3*C*), whereas miR-149-3p and miR-4488 were detected in 112 (57.4%) and 57 (30.8%) samples, respectively.

Then, miRNA levels in FIT cohort samples were explored by stratifying participants according to the colonoscopy results. Comparing the levels of the 25 stool DEmiRNAs between 46 participants with a negative colonoscopy result (excluding 7 participants with high hemoglobin levels) and 22 patients with CRC, 8 (let-7a-5p, let-7i-5p, miR-148a-3p, let-7b-5p, miR-320a-3p, miR-12114, miR-21-5p, miR-607-5p) were significantly different (adjusted $P < .05$) (Supplementary Table 4*E* and Figure 3*C*). Correlating the miRNA levels in FIT leftovers with the hemoglobin levels, only let-7b-5p showed a significant but limited correlation (rho = 0.16, $P < .05$) (Supplementary Table 4*F*).

Interestingly, miR-1246 and miR-607-5p were characterized, respectively, by increasing and decreasing levels, from colonoscopy-negative participants to CRC patients, as observed in the stool of the 3 case-control cohorts initially investigated for the miRNA signature identification (Figure 3*D*).

Comparable miRNA expression levels and variability were observed between paired FIT leftover/stool samples from 57 individuals analyzed by small RNA-seq (rho = 0.70, $P < .001$) (Supplementary Table 1*B* and Supplementary Figure 2*C*). Considering the levels of 468 miRNAs detected in at least half of FIT leftover samples, 99.6% were coherent with those in stool, with 282 miRNAs significantly correlated (average rho = 0.39, $P < .05$) (Figure 3*C*, Supplementary Figure 2*C*, and Supplementary Table 4*D*). In both sample types, miR-3125-3p, miR-6075-5p, and miR-1246 were characterized by the highest levels, and miR-3125-3p was detected in all samples and associated with the lowest expression variability, in agreement with our previous findings in stool samples of 335 control individuals[25] (Supplementary Figure 3*A* and Supplementary Table 4*D*). The levels of all 25 stool DEmiRNAs positively correlated between the 2 specimens, with 13 of them reaching statistical significance (including miR-607-5p, miR-1246, miR-149-3p, and miR-4488 from the 5-miRNA signature; $P < .05$) (Figure 3*C* and Supplementary Table 4*D*).

The 5-miRNA signature analyzed in FIT buffer leftovers was finally tested for the classification of patients with CRC from control individuals considering the signature alone or in combination with patient age, sex, and FIT hemoglobin levels. The 5-miRNA signature alone showed comparable classification performance (AUC, 0.85) as using age, sex, and hemoglobin levels (AUC, 0.87), and the combination of both data provided the best classification results (AUC, 0.93) (Supplementary Table 3*D*).

Discussion

In the present study, to our knowledge, we performed the first large-scale profiling of the stool miRNome by deep sequencing of samples from patients with CRC, colorectal polyps, or other GI diseases and control individuals. Given the pervasive detection across multiple cohorts, we confirmed previous findings about fecal miRNA potential use as noninvasive molecular biomarkers[23] (Supplementary Table 1*C* and Supplementary Figure 3*A*). We also reported novel evidence on specific markers across different disease conditions. Notably, a fecal miRNA signature was able to accurately distinguish CRC patients from control individuals: both its ability to distinguish AA and its detection in FIT leftovers support future investigations for a use in CRC screening implementation.

In CRC patients, 25 fecal miRNAs emerged coherently altered in 2 independent cohorts. The profile of these miRNAs in stool reflected their altered expression in tumor tissue or adjacent colonic mucosa. More than half of such DEmiRNAs were already reported as altered in CRC, either in tissue or in various biofluids, including the up-regulated miR-21-5p, miR-148a-3p, miR-149-3p, miR-194-5p, miR-200b-3p, and miR-320a-3p (Supplementary Table 5*A*).[23,36] Other miRNAs were associated with a disease for the first time by us; thus, further in vitro studies are needed to characterize the functional activity of these molecules and their involvement in CRC. Moreover, 3 DEmiRNAs identified in our study (miR-4323-5p, miR-607-5p, and miR-922-5p) are not currently annotated in the miRbase but were quantified based on the read mapping position within the miRNA hairpin. This is consistent with the need for

Figure 2. (*A*) Scatterplot reporting the stool miRNA average levels in CRC patients (*y*-axis) or control individuals (*x*-axis) from the IT cohort (*left*) or CZ cohort (*right*). The dot color represents the log2 fold change (log2FC) from the differential expression analyses between CRC and healthy individuals, and the size is proportional to the age, sex, and multiple-testing adjusted *P* values. (*B*) Scatterplot reporting the correlations of log2FC of the 25 DEmiRNAs from the comparison between CRC and control individuals and in common between the IT cohort (*x*-axis) and the CZ cohort (*y*-axis). The up-regulated and down-regulated miRNAs are reported in red and blue, respectively. (*C*) Heatmap of stool DEmiRNA levels in CRC and control individuals of both cohorts. For each participant, the CRC stage and grade based on the American Joint Committee on Cancer system, presence of metastasis, lymph node invasion status (pN), tumor size (pT), tumor localization, cohort of origin, and disease status (CRC or control) are reported. (*D*) DEmiRNA levels comparing CRC patients stratified for clinical data. The dot color represents the log2FC, and the dot size is proportional to the statistical significance. Black borders represent tests with $P < .05$. (*E*) Line plot reporting the ability of different combinations of feature selection methods and classifiers to perform the classification of CRC and control individuals. Each dot represents an AUC obtained using a different number of fecal DEmiRNAs in input. (*F*) Receiver operating characteristic curves obtained for the classification of CRC and control individuals using the identified miRNA signature. Data are reported for the 30% of participants excluded from the training set (*left*) and for the validation cohort (*right*). Adj., adjusted.

Table 2. Expression Levels and Fold Changes of the 25 Stool DEmiRNAs in Common Between the IT and CZ Cohorts

				Median levels, controls		Median levels, CRC		log2FC		Benjamini-Hochberg adjusted *P* value[a]	
ID	miRNA gene ID	Chromosome	Genomic context	IT cohort	CZ cohort	IT cohort	CZ cohort	IT cohort	CZ cohort	IT cohort	CZ cohort
let-7a-5p	MIRLET7A3	chr22	Intergenic	52.18	28.12	717.25	50.53	5.44	1.39	2.51E–24	1.05E–02
let-7b-5p	MIRLET7B	chr22	Intergenic	20.19	12.94	474.50	26.28	4.63	1.83	3.04E–19	6.54E–03
let-7f-5p	MIRLET7F1/MIRLET7F2	chr9/chrX	Intergenic/intron (*HUWE1*)	54.93	33.83	513.72	38.72	5.41	1.40	2.27E–27	1.05E–02
let-7i-5p	MIRLET7I	chr12	Partial overlap (*LINC01465*)	16.75	10.68	577.93	27.38	5.68	2.49	1.25E–23	6.54E–04
miR-1181	MIR1181	chr19	Exon (*CDC37*)	72.46	38.12	83.60	65.61	0.64	0.78	1.12E–02	4.63E–02
miR-12114	MIR12114	chr22	Intron (*PPP6R2*)	126.48	43.52	266.97	67.67	1.50	1.53	1.06E–07	4.71E–03
miR-1246	MIR1246	chr2	Intron (*LINC01117*)	909.33	568.34	2970.91	2364.91	3.59	2.83	9.63E–17	3.98E–06
miR-1290	MIR1290	chr1	Intron (*ALDH4A1*)	46.70	33.77	231.36	82.25	3.73	2.29	1.71E–21	4.13E–04
miR-148a-3p	MIR148A	chr7	Intergenic	19.17	11.56	425.27	25.82	5.60	2.27	4.19E–22	1.92E–03
miR-149-3p	MIR149	chr2	Intron (*GPC1*)	30.82	16.15	34.55	36.97	0.58	0.96	1.89E–02	3.92E–02
miR-194-5p	MIR194-1 / MIR194-2	chr1 / chr11	Intron (*IARS2*)/intergenic	69.85	59.45	206.31	68.59	3.63	1.02	3.44E–20	2.38E–02
miR-200b-3p	MIR200B	chr1	Intergenic	22.03	20.39	204.93	23.29	5.16	1.43	2.85E–23	2.01E–02
miR-21-5p	MIR21	chr17	Exon (*VMP1*)	37.68	42.23	557.19	63.56	5.36	1.78	1.15E–22	1.22E–02
miR-26a-5p	MIR26A1 / MIR26A2	chr3 / chr12	Intron (*CTDSPL*)/intron (*CTDSPL2*)	36.78	33.23	425.88	44.01	4.77	1.59	2.85E–23	1.68E–02
miR-320a-3p	MIR320A	chr8	Intergenic	27.26	16.26	271.19	33.93	3.29	1.50	1.01E–15	5.33E–03
miR-4323-5p	MIR4323	chr19	Intron (POU2F2-AS1)	67.11	29.50	73.39	58.96	1.62	1.92	8.88E–07	5.12E–03
miR-4488	MIR4499	chr11	Intergenic	113.12	50.73	342.90	73.67	2.53	1.23	2.94E–19	2.91E–02
miR-4492	MIR4492	chr11	Exon/intron (*BCL9L*)	25.04	14.50	34.76	22.24	1.28	1.26	1.62E–06	7.47E–03
miR-4508	MIR4508	chr15	Intergenic	94.44	34.09	98.33	86.36	0.87	1.12	3.85E–04	2.56E–02
miR-607-5p	MIR607	chr10	Intergenic	222.53	132.30	51.44	87.13	–1.72	–0.88	2.17E–18	6.54E–03
miR-6076	MIR6076	chr14	Intron (*LINC01588*)	32.14	23.14	15.10	15.54	–0.68	–1.24	1.05E–02	1.83E–02
miR-6131	MIR6131	chr5	Intergenic	31.05	15.50	103.66	22.39	2.08	1.49	2.19E–12	3.31E–03
miR-6777-5p	MIR6777	chr17	Intron (*SREBF1*)	235.14	140.02	42.53	80.22	–1.60	–1.02	4.60E–08	1.29E–02
miR-922-5p	MIR922	chr3	Exon (*RUBCN*)	335.74	206.43	71.51	89.57	–2.06	–1.26	1.99E–11	3.92E–02
miR-9899	MIR9899	chr2	Intron (*LYPD6*)	71.25	50.86	33.99	26.40	–0.55	–1.03	1.09E–02	4.00E–02

chr, chromosome; ID, identifier; log2FC, log2 fold change.

[a] Age- and sex-adjusted analysis.

Table 3. Performance of the 5-miRNA Predictive Signature in the Different Comparisons

Analysis details[a]							Precision		F1 score	
Comparison	Validation set	AUC (Mean ± SD)	95% CI	Accuracy	Sensitivity	Specificity	Disease	Control	Disease	Control
CRC vs control individuals	IT cohort + CZ cohort[b]	0.86 ± 0.01	0.79–0.94	0.78	0.78	0.78	0.82	0.74	0.80	0.76
CRC vs control individuals	Validation cohort	0.96 ± 0.01	0.92–1.00	0.89	0.90	0.88	0.93	0.83	0.91	0.85
Stage I–II CRC vs control individuals	IT cohort + CZ cohort[b]	0.86 ± 0.01	0.76–0.96	0.81	0.65	0.90	0.79	0.82	0.71	0.86
Stage I–II CRC vs control individuals	Validation cohort	0.95 ± 0.01	0.90–1.00	0.86	0.82	0.91	0.90	0.83	0.86	0.87
Stage III–IV CRC vs control individuals	IT cohort + CZ cohort[b]	0.88 ± 0.01	0.78–0.98	0.83	0.66	0.92	0.82	0.83	0.73	0.88
Stage III–IV CRC vs control individuals	Validation cohort	0.96 ± 0.01	0.91–1.00	0.85	0.75	0.94	0.91	0.82	0.82	0.88
CRC + AA vs control individuals	IT cohort + CZ cohort[b]	0.84 ± 0.01	0.77–0.91	0.77	0.83	0.67	0.81	0.70	0.81	0.69
AA vs control individuals	IT cohort + CZ cohort[b]	0.82 ± 0.01	0.71–0.97	0.79	0.61	0.86	0.62	0.85	0.62	0.85
AA + nAA vs control individuals	IT cohort + CZ cohort[b]	0.77 ± 0.02	0.65–0.89	0.73	0.62	0.81	0.67	0.77	0.64	0.79
nAA vs control individuals	IT cohort + CZ cohort[b]	0.80 ± 0.01	0.63–0.97	0.82	0.13	0.99	0.79	0.82	0.22	0.90
CRC vs AA	IT cohort + CZ cohort[b]	0.68 ± 0.02	0.54–0.82	0.76	0.92	0.25	0.80	0.49	0.85	0.33

[a] Analysis includes age and sex covariates.
[b] Thirty percent of samples were excluded from the training and matched by age, sex, cohort, and CRC stage.

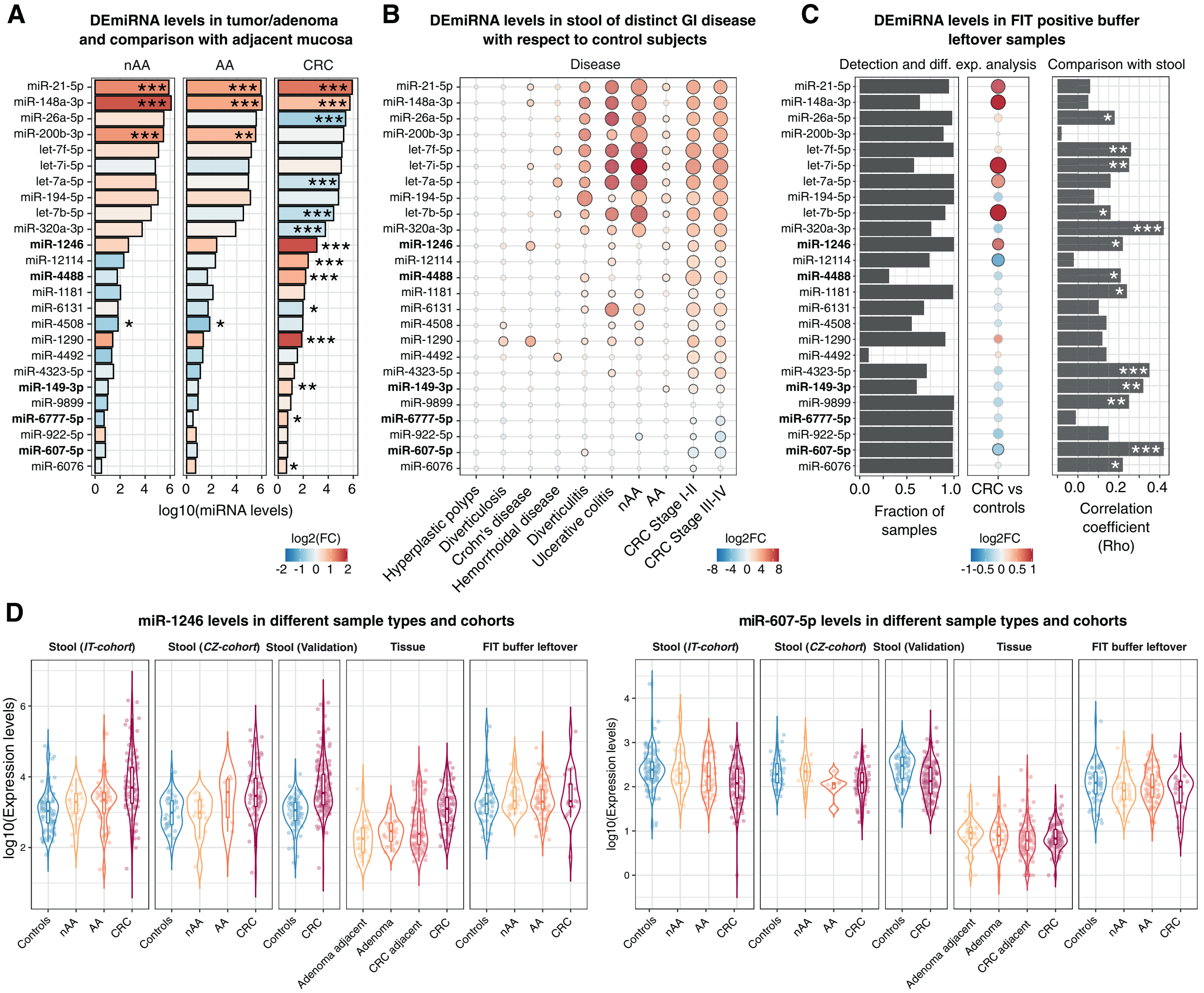


Figure 3. Characterization of the 25 fecal DEmiRNAs in different sample types. (*A*) Bar plot reporting the median levels in tumor, AA, and nAA tissues. The color code represents the log2 fold change (log2FC) from the paired differential expression analysis between CRC/adenoma tissues and matched adjacent mucosa. ***Adjusted $P < .001$, **adjusted $P < .01$, *adjusted $P < .05$. (*B*) Comparison of miRNA levels in the stool of patients with CRC, colorectal adenomas, hyperplastic polyps, or other GI disorders with respect to control individuals. The dot color represents the log2FC, and the dot size is proportional to the analysis significance. Black borders represent results with an adjusted $P < .05$. (*C*) DEmiRNA analysis in FIT leftover samples from CRC screening. (*Left*) The fraction of FIT cohort samples in which each miRNA was detected and (*center*) results of the differential expression analysis between FIT-positive patients with CRC diagnosis based on colonoscopy outcome and those with a negative one. The dot color represents the log2FC, and the dot size is proportional to the analysis significance. Black borders represent a DESeq2 Benjamini-Hochberg adjusted $P < .05$. (*Right*) Correlation coefficients between miRNA levels in stool and FIT buffer leftover samples from the same individuals (***$P < .001$, *$P < .05$). (*D*) Box plots reporting miR-1246 and miR-607-5p levels in all study cohorts and biospecimens.

continuous refinement of miRBase annotations[37] and with evidence of new miRNAs reported by different groups.[38,39]

Consistent with their overall higher/lower levels in the stool of CRC patients with respect to that of control individuals, the 25 DEmiRNA levels also increased/decreased with tumor size and stage. On the other hand, they were characterized by coherent altered levels when patients were stratified by tumor localization (proximal, distal, rectum) (Supplementary Table 4*C*). This further supports the importance of these miRNAs in relationship with the disease, as confirmed by the overrepresentation of cancer-related processes involving their validated target genes (Supplementary Table 2*D* and *E*).

Based on this initial evidence, we implemented an integrated explainable ML strategy to explore, among the 25 DEmiRNAs, the minimal set of stool miRNAs able to accurately discriminate CRC patients from control individuals. Our approach generated a signature composed of 5 miRNAs (namely, miR-1246, miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p) that was clinically validated in an additional independent cohort of cases compared to healthy volunteers and technically validated by another methodology (ie, RT-qPCR). The accurate discrimination of both participants in early and late cancer stages from control individuals confirmed the robustness of these 5 miRNAs for CRC detection. Although based on a small sample set, the signature could also accurately discriminate participants with AA from control individuals (AUC, 0.86), and in all analyses, high performances were obtained, irrespectively, by adjusting or not for sex and age, 2 relevant risk factors for this cancer.[40] To the best of our knowledge, this is the first signature based on fecal miRNAs whose efficiency was proven in populations from 2 countries characterized by different lifestyle and dietary habits[41] and CRC incidence.[42] Notably, such populations also show different trends in early-onset CRC,[43] the incidence of which is linked to unhealthy individual habits, such as a sedentary lifestyle.[44]

Similar to the functional analysis of all 25 DEmiRNAs, focused research on the 5-signature miRNA target genes evidenced a prevalence of genes involved in cancer-related processes, including regulation of the cell cycle, programmed cell death, and DNA damage response. Interestingly, functional analysis of predicted target genes of miR-607-5p highlighted terms/processes related to nuclear cell cycle DNA replication and showed *TRIM66*, *HIPK2*, *GRIN2B*, and *WTIP* as the targets with the highest number of miR-607-5p binding sites (Supplementary Table 5*B* and *C*).

Among all the miRNAs of the signature, miR-1246 has been previously widely studied in CRC. Altered levels of this miRNA have been found in circulating exosomes in relation to cancer metastasis and prognosis.[45,46] Exosomal miR-1246 levels were induced by *Fusobacterium nucleatum* in in vitro and in vivo CRC models with an increase of tumor cell metastatic potential.[47] These results align with more recent observations on the relationship between intratumor levels of *F nucleatum* and the aggressiveness of colon and breast cancers.[48] An intratumor increase in this well-known CRC-related bacteria might induce the release of exosomal miR-1246 in the gut lumen, with the subsequent detection of this miRNA in stool samples. Similar considerations could be drawn from another study investigating a model of enterotoxigenic *Bacteroides fragilis* that induced up-regulation of exosomal miR-1246 in CRC cell lines.[49] Interestingly, in the same study, this microbial species reduced the exosomal levels of another fecal miRNA included in our signature, miR-149-3p, that was demonstrated to regulate tumor-infiltrating CD4$^+$ T-helper type 17 differentiation.[49]

Similar findings were observed when analyzing the fecal miRNome and gut metagenome data from a previous study by our group in which we investigated the miRNA-microbiota relationships in stool samples.[20] Specifically, by reanalyzing the data from that study, miR-1246 levels emerged as significantly related to both *F nucleatum* and *B fragilis* abundances, whereas miR-149-3p was inversely related to *B fragilis* abundances (Supplementary Figure 3*B*). This pervasive relationship between in vitro exosomal miRNA levels and microbial infections suggests that the most informative stool biomarkers for CRC might reflect the dysregulated interactions between colonic tissue and the gut microbiota. Interestingly, in the miRNA-microbiota correlation analysis, 2 down-regulated fecal miRNAs (miR-607-5p and miR-6777-5p), included in the predictive signature and so far scantly investigated in the literature, were inversely related not only to *F nucleatum* and *B fragilis* abundances but also to *Escherichia coli*, another species related to CRC onset[50] (Supplementary Figure 3*B*).

To further explore the stool results, we tested DEmiRNA patterns in tumor and adenoma tissues paired with nonmalignant adjacent mucosa from patients of the IT cohort. Stool generally mirrored the altered miRNA expression levels of these tissues. Only the levels of miR-21-5p and miR-148a-3p increased in both CRC and adenoma compared to matched adjacent mucosa, whereas the other DEmiRNAs (including miR-1246, miR-4488, and miR-149-3p of the signature) showed a CRC-specific dysregulation. miR-607-5p and miR-6777-5p, decreasing in patients' fecal samples, were characterized by low expression levels in both tumor/adenoma and adjacent mucosa, suggesting their deletion or epigenetic silencing. In The Cancer Genome Atlas,[51] both miRNAs are frequently deleted in CRC (Supplementary Table 5*D*), supporting the down-regulation in stool and tumor tissues observed by us. In agreement with our findings, previous studies have demonstrated that the down-regulation of miRNAs seems to be a premature step in the development of several cancers.[52,53] Surprisingly, miR-320a, let-7b-5p, and let-7a-3p, more abundant in stool of CRC patients, were more expressed in adjacent mucosa than in tumor tissue. miR-320a has been widely reported as down-regulated in CRC,[54] whereas its circulating levels increased in relation to gut inflammation in IBD patients,[55] coherent with our data in stool samples. Interestingly, miR-320a has been described as a key regulator of intestinal barrier formation.[56] Similarly, the expression of let-7 family members has been observed in the healthy gut epithelium, whereas their genetic depletion induced tumorigenesis in CRC mouse models.[57] Thus, the analysis of stool miRNAs is relevant to identify not only markers of the tumor small noncoding transcriptome but may also unveil an intestinal response of the stromal component to the presence of a tumor mass.

We also explored the miRNome of plasma EVs from a subset of the study population using the same experimental approach as in stool and tissue samples. However, in this circulating biospecimen, only a few miRNAs showed similar trends as in feces. For instance, among the miRNAs of the signature, miR-1246 and miR-4488 levels significantly increased in plasma EVs of CRC patients compared with control individuals. These results are consistent with previous findings reported by us, supporting stool miRNAs as more sensitive than plasma miRNAs in reflecting intestinal changes driven by a long-term dietary pattern.[24] Although more data are needed to compare the stool and plasma EV miRNome, given the reported relationships between miR-1246 levels in EVs and CRC metastasis,[45] these circulating molecules may be more informative for advanced stages of the disease, which is beyond the scope of our investigation.

In this study, we sought to compare the stool DEmiRNA profiles of CRC patients with those of patients with other bowel inflammatory diseases of different severity confirmed by colonoscopy. Besides different polyp types, we included samples from several GI diseases, like different types of IBDs and diverticulitis. Notably, although the CRC-specific miRNAs were down-regulated, most of the altered miRNAs in common with adenomas and inflammatory diseases were up-regulated: miR-21-5p was the clearest example, confirming the literature.[26] As an exception, miR-607-5p was down-regulated in the stool miRNA profiles of patients with AA and ulcerative colitis. Accordingly, recent studies showed altered miRNA profiles in the fecal samples of patients with inflammation,[58,59] even in relation to microbiota.[60] We can therefore conclude that altered stool miRNA profiles reflect either the intestinal response to an inflammatory process or the transcriptional alterations related specifically to CRC development. Importantly, we clearly demonstrated that well-known CRC-related miRNAs, such as miR-21-5p, show dysregulated fecal levels in several disease contexts, suggesting that other miRNAs, such as miR-6777-5p and miR-149-3p, should be investigated to design CRC-specific molecular signatures. This is the first evidence from a large-scale analysis of individuals with different gastrointestinal diseases of stool miRNAs specifically altered in CRC. It also highlights an extensive reflection of the gut inflammation on the fecal miRNA levels.

The fact that specific dysregulated fecal miRNAs could distinguish individuals with CRC or precursor lesions from control individuals and that, at least for cancer, data were confirmed in different cohorts, encouraging their use to complement the existing noninvasive screening tests. In this respect, we also investigated whether miRNAs can be detected in buffer-diluted stool leftovers from FIT tubes used in a context of a population-based screening program, and we found a remarkable similarity between the profiles detected in the stool samples collected in nucleic acid preservative medium tubes from the same participants. Despite data on a larger cohort being needed, this pilot small RNA-seq–based quantification of miRNAs in FIT buffer leftovers is consistent with previous evidence measuring miRNAs in this sample type by RT-qPCR,[22] as well as by us. By exploring miRNA profiles within FIT-positive patients, we observed a subset of miRNAs differentially expressed between individuals with a positive or a negative colonoscopy outcome. In addition, miR-1246 and miR-607-5p from the 5-miRNA signature deserve further investigation because they were detected in most of the samples, and their levels respectively increased and decreased progressively, going from individuals with negative colonoscopy results, to those with adenomas of different severity, to CRC patients. Although these data confirm that miRNAs can be widely detected in FIT leftovers, the comparative results between individuals must be carefully considered given the small group size analyzed so far; the lack of samples from FIT-negative individuals; and the fact that we cannot rule out the role of confounding factors, including subclinical diseases in the colonoscopy-negative patients.

Most likely, by including hemoglobin levels evaluated by FIT, the discrimination capability of the present stool miRNA predictive signature would be further improved, as already reported in the past (FIT/FOBT + microbiome,[11,61] FIT + miRNAs,[21] and FIT + methylation markers[62]). The sensitivity and specificity of our 5-miRNA signature suggest that it could show a similar diagnostic performance as the multitarget stool DNA test[63] when used as a screening test in average-risk populations. Duran-Sanchon et al[21] proposed a 2-stool miRNA-based classification signature (namely, miR-27a-3p and miR-421) combined with hemoglobin levels, age, and sex of FIT-positive individuals. The signature accurately classified CRC (AUC, 0.93) from control individuals but was less efficient when AA patients were included (AUC, 0.70).[62] Different from us, the researchers initially selected miRNAs based on their differential expression between tumor tissue and adjacent mucosa and included in all models sex and age, 2 important risk factors for CRC. Hereby, we demonstrated the robustness of our signature because its performance remained similar even without the inclusion of age and sex covariates. In addition, despite the study not being designed for identifying stool biomarkers for adenomas, the 5-miRNA signature was able to accurately distinguish AA alone or in combination with CRC (AUC, 0.84), suggesting its use to detect precancer lesions at risk. In our study, miR-27a-3p and miR-421 were detected in tissue samples but not in stool, where only the former miRNA was measurable. In search of reproducible fecal molecular biomarkers for the noninvasive diagnosis of CRC and adenomas,[11] a hypothesis-free miRNome-wide approach, such as the small RNA-seq analysis in stool performed in multiple independent populations, overcomes these issues.

The present study has several strengths: (1) the inclusion of independent cohorts from 2 countries with different diet and lifestyle habits as well as CRC rates; (2) the fact that the cohorts were different for CRC clinical characteristics, allowing the identification of accurate biomarkers independent of the disease stage; (3) the adoption of the same protocol for the collection of stool in both training cohorts; (4) the validation of the signature on a cohort with a different stool collection protocol, showing its robustness; (5) the miRNome-wide approach in different biospecimens and different GI disease contexts, which has allowed us to discriminate miRNAs specifically dysregulated in the stool of CRC patients; 6) the implementation of an explainable ML approach able to provide an unbiased method for identifying the minimal set of predictive biomarkers.

However, we are also aware of several limitations. Although there was a similar study design for recruitment, the 2 cohorts were heterogeneous for individual cancer categories. This heterogeneity could be responsible for the observed differences in the median stool miRNA levels and expression differences between the 2 cohorts. Given the difference in the clinical characteristics of CRC patients, the main driver of such a difference may be the higher proportion of low-grade and low-stage tumors in the CZ cohort. However, the fact that the results are reproducible between cohorts further supports the robustness of the signature identified in this study.

Despite the large number of analyzed samples, the variegated spectrum of CRC, adenomas, and other precancerous lesions needs to be more exhaustively represented and deserves further investigation. For example, we did not investigate serrated lesions or deeply explore the alterations in CRC stratified based on molecular or clinical data. In addition, even though the observed DEmiRNAs were not reported to be modulated by dietary habits,[24] the lack of dietary/lifestyle information of analyzed individuals may represent a limitation of the study. Follow-up studies with additional cohorts representing patients with different ethnicities, dietary patterns, and lifestyle habits are required, but this is beyond the scope of this study, which, to our knowledge, represents the largest sequencing-based analysis of stool miRNAs so far.

In conclusion, this multicenter and international study based on small RNA-seq allowed us to comprehensively detect in stool several miRNAs differentially expressed in CRC. Furthermore, the implemented ML approach identified a minimal number of miRNAs whose combined profiles showed a good discriminating power for the presence of a tumor or AA, independent of age and sex. This may represent a fecal signature for improving the effectiveness of current noninvasive screening programs, potentially increasing sensitivity and maintaining high specificity, and applicable on a large scale, with a reasonable cost/time required.

In this respect, for FIT implementation, in the near future miRNA profiles will be investigated in additional cohorts, possibly from different countries, increasing the number/types of precancer lesions and also including FIT-negative samples, with the chance to explore the role of diet and lifestyle habits on an adequate scale. Furthermore, the inclusion of FIT-negative samples will allow the possibility to prospectively test miRNA profiles in subsequent rounds of CRC screening, collecting multiple samples per individual. In parallel, the analysis of the microbiome composition of stool/leftover FIT samples will help deepen the research on gut-host crosstalk with small noncoding RNAs. Finally, even if small RNA-seq and RT-qPCR currently represent the most commonly used approaches for miRNA analyses, we must consider that more rapid, practical, but reliable approaches, such as biosensors, may provide an alternative for testing the miRNA signature in a large clinical setting.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2023.05.037.

References

1. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol 2019;16:713–732.
2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249.
3. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144:1941–1953.
4. Kral J, Kojecky V, Stepan M, et al. The experience with colorectal cancer screening in the Czech Republic: the detection at earlier stages and improved clinical outcomes. Public Health 2020;185:153–158.
5. Lauby-Secretan B, Vilahur N, Bianchini F, et al. The IARC perspective on colorectal cancer screening. N Engl J Med 2018;378:1734–1740.
6. Senore C, Basu P, Anttila A, et al. Performance of colorectal cancer screening in the European Union member states: data from the second European screening report. Gut 2019;68:1232–1244.
7. Rabeneck L, Chiu HM, Senore C. International perspective on the burden of colorectal cancer and public health effects. Gastroenterology 2020;158:447–452.
8. **Robertson DJ, Lee JK**, Boland CR, et al. Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2017;152:1217–1237.
9. Loktionov A. Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? World J Gastrointest Oncol 2020;12:124–148.
10. **Weng M, Wu D, Yang C**, et al. Noncoding RNAs in the development, diagnosis, and prognosis of colorectal cancer. Transl Res 2017;181:108–120.
11. **Thomas AM, Manghi P**, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–678.
12. **Sun Y, Guo Z**, Liu X, et al. Noninvasive urinary protein signatures associated with colorectal cancer diagnosis and metastasis. Nat Commun 2022;13(1):2757.
13. Francavilla A, Turoczi S, Tarallo S, et al. Exosomal microRNAs and other non-coding RNAs as colorectal cancer biomarkers: a review. Mutagenesis 2020;35:243–260.
14. Hombach S, Kretz M. Non-coding RNAs: classification, biology and functioning. Adv Exp Med Biol 2016;937:3–17.
15. Di Leva G, Croce CM. miRNA profiling of cancer. Curr Opin Genet Dev 2013;23:3–11.
16. Moridikia A, Mirzaei H, Sahebkar A, et al. MicroRNAs: potential candidates for diagnosis and treatment of colorectal cancer. J Cell Physiol 2018;233:901–913.
17. Dragomir MP, Kopetz S, Ajani JA, et al. Non-coding RNAs in GI cancers: from cancer hallmarks to clinical utility. Gut 2020;69:748–763.
18. Pardini B, Sabo AA, Birolo G, et al. Noncoding RNAs in extracellular fluids as cancer biomarkers: the new frontier of liquid biopsies. Cancers (Basel) 2019;11(8):1170.
19. Cervena K, Novosadova V, Pardini B, et al. Analysis of MicroRNA expression changes during the course of therapy in rectal cancer patients. Front Oncol 2021;11:702258.
20. **Tarallo S, Ferrero G**, Gallo G, et al. Altered fecal small RNA profiles in colorectal cancer reflect gut microbiome composition in stool samples. mSystems 2019;4(5):e00289-19.
21. Duran-Sanchon S, Moreno L, Auge JM, et al. Identification and validation of microRNA profiles in fecal samples for detection of colorectal cancer. Gastroenterology 2020;158:947–957.
22. Zhao Z, Zhu A, Bhardwaj M, et al. Fecal microRNAs, fecal microRNA panels, or combinations of fecal microRNAs with fecal hemoglobin for early detection of colorectal cancer and its precursors: a systematic review. Cancers (Basel) 2021;14(1):65.

23. Francavilla A, Tarallo S, Pardini B, et al. Fecal microRNAs as non-invasive biomarkers for the detection of colorectal cancer: a systematic review. Minerva Biotecnol 2019;31:30–42.
24. **Tarallo S, Ferrero G, De Filippis F**, et al. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. Gut 2021;71:1302–1314.
25. **Francavilla A, Gagliardi A, Piaggeschi G**, et al. Faecal miRNA profiles associated with age, sex, BMI, and lifestyle habits in healthy individuals. Sci Rep 2021;11(1):20645.
26. Jenike AE, Halushka MK. miR-21: a non-specific biomarker of all maladies. Biomark Res 2021;9(1):18.
27. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501–1504.
28. **Wirbel J, Pyl PT**, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–689.
29. Lin Y, Lau HC, Liu Y, et al. Altered mycobiota signatures and enriched pathogenic *Aspergillus rambellii* are associated with colorectal cancer based on multicohort fecal metagenomic analyses. Gastroenterology 2022;163:908–921.
30. Zwinsová B, Petrov VA, Hrivňáková M, et al. Colorectal tumour mucosa microbiome is enriched in oral pathogens and defines three subtypes that correlate with markers of tumour progression. Cancers (Basel) 2021; 13(19):4799.
31. **Francavilla A, Ferrero G, Pardini B**, et al. Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk. Gut Microbes 2023;15(1): 2172955.
32. **Sabo AA, Birolo G, Naccarati A**, et al. Small non-coding RNA profiling in plasma extracellular vesicles of bladder cancer patients by next-generation sequencing: expression levels of miR-126-3p and piR-5936 increase with higher histologic grades. Cancers (Basel) 2020;12(6):1507.
33. **Moisoiu T, Dragomir MP, Iancu SD**, et al. Combined miRNA and SERS urine liquid biopsy for the point-of-care diagnosis and molecular stratification of bladder cancer. Mol Med 2022;28(1):39.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
35. Zhang J, Storey KB. RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. PeerJ 2018;6:e4262.
36. Slaby O. Non-coding RNAs as biomarkers for colorectal cancer screening and early detection. Adv Exp Med Biol 2016;937:153–170.
37. **Alles J, Fehlmann T**, Fischer U, et al. An estimate of the total number of true human miRNAs. Nucleic Acids Res 2019;47:3353–3364.
38. **Jima DD, Zhang J**, Jacobs C, et al. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. Blood 2010;116:e118–e127.
39. **Friedlander MR, Lizano E, Houben AJ**, et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome Biol 2014;15(4):R57.
40. Wei EK, Giovannucci E, Wu K, et al. Comparison of risk factors for colon and rectal cancer. Int J Cancer 2004; 108:433–442.
41. Imamura F, Micha R, Khatibzadeh S, et al. Dietary quality among men and women in 187 countries in 1990 and 2010: a systematic assessment. Lancet Glob Health 2015;3(3):e132–e142.
42. **Wong MCS, Huang J, Lok V**, et al. Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. Clin Gastroenterol Hepatol 2021;19:955–966.
43. Vuik FE, Nieuwenburg SA, Bardou M, et al. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. Gut 2019;68: 1820–1826.
44. Patel SG, Karlitz JJ, Yen T, et al. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. Lancet Gastroenterol Hepatol 2022;7:262–274.
45. Desmond BJ, Dennett ER, Danielson KM. Circulating extracellular vesicle microRNA as diagnostic biomarkers in early colorectal cancer—a review. Cancers (Basel) 2019;12(1):52.
46. Cooks T, Pateras IS, Jenkins LM, et al. Mutant p53 cancers reprogram macrophages to tumor supporting macrophages via exosomal miR-1246. Nat Commun 2018;9(1):771.
47. Guo S, Chen J, Chen F, et al. Exosomes derived from *Fusobacterium nucleatum*-infected colorectal cancer cells facilitate tumour metastasis by selectively carrying miR-1246/92b-3p/27a-3p and CXCL16. Gut 2021; 70:1507–1519.
48. **Fu A, Yao B, Dong T**, et al. Emerging roles of intratumor microbiota in cancer metastasis. Trends Cell Biol 2023; 33:583–593.
49. **Cao Y, Wang Z, Yan Y**, et al. Enterotoxigenic *Bacteroides fragilis* promotes intestinal inflammation and malignancy by inhibiting exosome-packaged miR-149-3p. Gastroenterology 2021;161:1552–1566.
50. Clay SL, Fonseca-Pereira D, Garrett WS. Colorectal cancer: the facts in the case of the microbiota. J Clin Invest 2022;132(4):e155101.
51. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487(7407):330–337.
52. Esquela-Kerscher A, Slack FJ. Oncomirs – microRNAs with a role in cancer. Nat Rev Cancer 2006;6:259–269.
53. **Vila-Navarro E, Vila-Casadesus M**, Moreira L, et al. MicroRNAs for detection of pancreatic neoplasia: biomarker discovery by next-generation sequencing and validation in 2 independent cohorts. Ann Surg 2017; 265:1226–1234.
54. Liang Y, Li S, Tang L. MicroRNA 320, an anti-oncogene target miRNA for cancer therapy. Biomedicines 2021; 9(6):591.
55. Cordes F, Demmig C, Bokemeyer A, et al. MicroRNA-320a monitors intestinal disease activity in patients with inflammatory bowel disease. Clin Transl Gastroenterol 2020;11(3):e00134.

56. Muenchau S, Deutsch R, de Castro IJ, et al. Hypoxic environment promotes barrier formation in human intestinal epithelial cells through regulation of microRNA 320a expression. Mol Cell Biol 2019;39(14):e00553-18.

57. Madison BB, Jeganathan AN, Mizuno R, et al. Let-7 represses carcinogenesis and a stem cell phenotype in the intestine via regulation of Hmga2. PLoS Genet 2015; 11(7):e1005408.

58. Wohnhaas CT, Schmid R, Rolser M, et al. Fecal microRNAs show promise as noninvasive Crohn's disease biomarkers. Crohns Colitis 360 2020;2(1):otaa003.

59. **Verdier J, Breunig IR**, Ohse MC, et al. Faecal microRNAs in inflammatory bowel diseases. J Crohns Colitis 2020;14:110–117.

60. **Ambrozkiewicz F, Karczmarski J**, Kulecka M, et al. In search for interplay between stool microRNAs, microbiota and short chain fatty acids in Crohn's disease—a preliminary study. BMC Gastroenterol 2020;20(1):307.

61. **Xie YH, Gao QY, Cai GX**, et al. Fecal *Clostridium symbiosum* for noninvasive detection of early and advanced colorectal cancer: test and validation studies. EBioMedicine 2017;25:32–40.

62. Bosch LJ, Oort FA, Neerincx M, et al. DNA methylation of phosphatase and actin regulator 3 detects colorectal cancer in stool and complements FIT. Cancer Prev Res (Phila) 2012;5(3):464–472.

63. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014;370:1287–1297.

Author names in bold designate shared co-first authorship.

Received August 8, 2022. Accepted May 24, 2023.

Correspondence
Address correspondence to: Alessio Naccarati, PhD, Italian Institute for Genomic Medicine, c/o IRCCS Candiolo, SP 142, Km 3.95, 10060 Candiolo, Turin, Italy. e-mail: alessio.naccarati@iigm.it; or Barbara Pardini, PhD, Italian Institute for Genomic Medicine, c/o IRCCS Candiolo, SP 142, Km 3.95, 10060 Candiolo, Turin, Italy. e-mail: barbara.pardini@iigm.it.

Acknowledgments
The authors are very grateful to all volunteers who participated with enthusiasm in the study. They are also thankful to Dr Vladimir Benes (Genecore Facility at EMBL, Heidelberg, Germany) for his kind collaboration. The artwork for this study has been created with BioRender.com.

CRediT Authorship Contributions
Barbara Pardini, PhD (Data curation: Equal; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Lead; Methodology: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Equal).

Giulio Ferrero, PhD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Lead; Investigation: Equal; Methodology: Lead; Software: Lead; Supervision: Supporting; Validation: Equal; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Sonia Tarallo, PhD (Conceptualization: Lead; Data curation: Lead; Formal analysis: Supporting; Investigation: Lead; Methodology: Lead; Validation: Lead; Visualization: Equal; Writing – original draft: Lead; Writing – review & editing: Equal).

Gaetano Gallo, MD, PhD (Conceptualization: Supporting; Data curation: Lead; Investigation: Supporting; Resources: Lead; Writing – review & editing: Equal).

Antonio Francavilla, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Methodology: Equal; Validation: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Nicola Licheri, MSc (Data curation: Equal; Formal analysis: Equal; Resources: Equal; Software: Equal; Writing – review & editing: Supporting).

Mario Trompetto, MD, PhD (Data curation: Equal; Investigation: Supporting; Methodology: Supporting; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Giuseppe Clerico, MD, PhD (Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Carlo Senore, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Formal analysis: Supporting; Methodology: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Sergio Peyre, MD (Funding acquisition: Lead; Investigation: Supporting; Methodology: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

Veronika Vymetalkova, PhD (Data curation: Equal; Formal analysis: Supporting; Funding acquisition: Supporting; Methodology: Equal; Project administration: Equal; Validation: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Ludmila Vodickova, PhD (Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Supporting; Project administration: Supporting; Validation: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Vaclav Liska, MD, PhD (Data curation: Supporting; Investigation: Supporting; Resources: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Ondrej Vycital, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Miroslav Levy, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Peter Macinga, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Tomas Hucl, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Eva Budinska, PhD (Data curation: Supporting; Investigation: Supporting; Resources: Equal; Validation: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Pavel Vodicka, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Supporting; Resources: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Francesca Cordero, PhD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Lead; Investigation: Equal; Methodology: Equal; Supervision: Supporting; Validation: Supporting; Visualization: Supporting; Writing – original draft: Lead; Writing – review & editing: Lead).

Alessio Naccarati, PhD (Conceptualization: Lead; Data curation: Lead; Formal analysis: Lead; Funding acquisition: Lead; Investigation: Supporting; Methodology: Equal; Project administration: Lead; Resources: Equal; Supervision: Lead; Validation: Equal; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Conflicts of interest
The authors disclose no conflicts.

Funding
This work was supported by the Italian Institute for Genomic Medicine and Compagnia di San Paolo Torino, Italy (to Alessio Naccarati, Barbara Pardini, and Sonia Tarallo), by Lega Italiana per La Lotta contro i Tumori (to Francesca Cordero, Barbara Pardini, and Alessio Naccarati), by the Grant Agency of the Czech Republic (17-16857S to Alessio Naccarati, Pavel Vodicka, Ludmila Vodickova, and Veronika Vymetalkova and 22-05942S to Veronika Vymetalkova, Ludmila Vodickova, Tomas Hucl, and Vaclav Liska), by the Grant Agency of the Ministry of Health of the Czech Republic (AZV NV18-03-00199 to Veronika Vymetalkova, Pavel Vodicka, Ludmila Vodickova, Tomas Hucl, and Peter Macinga), by the Fondazione CRT (grant no. 2017.2025 to Francesca Cordero), and by the COST action TRANSCOLONCAN (CA17118). Pavel Vodicka and Ludmila Vodickova acknowledge support from the National Operation Program (National Institute for Cancer Research LX22NPO 05120). Eva Budinska thanks the RECETOX Research Infrastructure (No. LM2023069) financed by the Ministry of Education, Youth and Sports, and the OP RDE (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17_043/0009632 and the CETOCOEN PLUS project No. CZ.02.1.01/0.0/0.0/15_003/0000469) for supportive background. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825410 (ONCOBIOME project to Alessio Naccarati, Barbara Pardini, Eva Budinska, Francesca Cordero, and Sonia Tarallo). The research leading to these results has received funding from AIRC under IG 2020 - ID 24882 (PI: Alessio Naccarati) to Alessio Naccarati; and IG 2019 - ID 23473 (PI: Carlo Senore) to Carlo Senore. This publication reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

Data Availability
All data relevant to the study are included in the article or in the Supplementary Material. Raw data are available upon request to the corresponding author.

Supplementary Methods

Stool Study Cohorts

Italian cohort. Stool specimens as well as clinical and demographic data were collected from 317 individuals recruited in a hospital-based study at 1 hospital in Vercelli, Italy (Table 1 and Figure 1*A*). Based on the results of a completed colonoscopy examination with adequate bowel preparation, participants were classified into (1) 89 CRC patients (individuals with newly diagnosed sporadic CRC); (2) 74 polyps patients, stratified as hyperplastic polyps (n = 6), nAA (n = 20), or AA (n = 48); (3) 49 patients with GI disease, such as IBD (including Crohn's disease and indeterminate or ulcerative colitis) or diverticular disease; and (4) 105 control individuals.

AAs were defined based on the presence of high-grade dysplasia, villous component, or lesion length of >1 cm as defined by Zarchy and Ershoff.[1] Of this cohort, 93 stool samples (from 29 CRC patients, 27 polyps, 13 individuals with a GI disease, and 24 colonoscopy-negative control individuals) were used and have been described previously.[2–4]

Czech cohort. Stool specimens as well as clinical and demographic data were collected from a cohort of 162 Czech individuals recruited in 2 hospitals in Prague and 1 in Plzen, Czech Republic (Table 1 and Figure 1*A*). Based on colonoscopy results, participants were divided into (1) 66 CRC patients; (2) 28 individuals with colorectal polyps, grouped as hyperplastic polyps (n = 9), nAA (n = 13), and AA (n = 6); (3) 32 patients with other GI disorders; and (4) 36 colonoscopy-negative control individuals.

In both studies, colonoscopy was recommended for 2 main reasons: (1) because of the recommendation of the family doctor for various reasons (age of the patient, complaints in the gut, etc) or (2) because the patient had a positive FIT result (ie, there was blood in the stool at the time of the test, and therefore the individual was invited to have a colonoscopy to further investigate the reason for blood in stool). In any case, individuals with major GI diseases other than cancer were considered apart from those control individuals with a negative colonoscopy finding.

Validation cohort. Stool specimens from 141 CRC patients recruited in the hospital in Brno, Czech Republic,[5] and 80 stool samples of healthy volunteers contributing to science[6] were included as an independent validation cohort. Stool specimens from 141 CRC patients were obtained at a hospital in Brno, Czech Republic: these individuals were previously described by Zwinsova et al[5] and here are sequenced for the first time for small RNA-seq.

Stool samples of healthy volunteers contributing to science are a part (about 20%) of the cohort described and sequenced for small RNA-seq by Tarallo et al[6] and Francavilla et al.[7] The healthy volunteers are derived from a subgroup of healthy individuals (no cancer, no precancer lesions) nested from the omnivorous group described by Tarallo et al.[6] and Francavilla et al.[7] Only individuals with age >30 years were considered for the analysis.

Fecal immunochemical test cohort. FIT leftover samples collected from 185 participants with a positive result from FIT analysis in the CRC screening for the general population of Piedmont Region (Italy) were added to the study. Based on the results of a completed colonoscopy examination with adequate bowel preparation, the individuals were classified as control individuals (n = 53) or individuals with AA (n = 80) or nAA (n = 30) and with CRC (n = 22). Among the 185 participants, 57 also provided stool samples before undergoing colonoscopy.

Colonoscopy was recommended because the patients had abnormal or positive FIT results (ie, there was blood in the stool at the time of the test), and therefore they were invited to have a colonoscopy to further investigate the reason for blood in stool.

Other Analyzed Biospecimens

For 132 patients (102 CRC patients and 30 patients with colorectal adenoma) primary CRC/adenoma tissues paired with adjacent colonic mucosa were collected in the same hospital as IT cohort. Among these patients, 69 (51 CRC and 18 colorectal adenoma) donated their stool and plasma samples and were included in the IT cohort.

Blood samples were collected from 210 participants of the IT cohort, stratified as 52 patients with CRC, 19 with AAs, 15 with nAAs, 6 with hyperplastic polyps, 34 with other GI disorders, and 79 control individuals.

Sample Collection

Naturally evacuated fecal samples were obtained from all participants previously instructed to self-collect the specimen at home. For all cohorts, stool samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp) and returned to the endoscopy unit. Stool aliquots (200 μL) were stored at –80°C until RNA extraction.[6,8] The only exception was represented by the validation cohort of CRC patients from Brno, for which stool samples were collected from untreated patients before the scheduled surgery with DNA-free swabs (Deltalab). Patients performed the collection at home before their hospitalization for the surgery and brought the samples to the hospital, where they were immediately frozen at –80°C until further processing.

For the FIT cohort, leftovers from FIT tubes ($\sim$1.2 mL) used for automated tests (OC-sensor, Eiken Chemical Co) for hemoglobin quantification were also collected and stored at –80°C until use.

Plasma samples were obtained from 8 mL of blood centrifuged for 10 minutes at 1000 revolutions/minute, and aliquots were stored at –80°C until use. Plasma exosomes/EVs were isolated from 200 μL of plasma using the ExoQuick exosome precipitation solution (System Biosciences, Mountain View), according to the manufacturer's instructions.[9,10] Briefly, plasma was mixed with 50.4 μL of ExoQuick solution and refrigerated at 4°C overnight (at least 12 hours). The mixture was then further centrifuged at 1500*g* for 30 minutes. The EV pellet was dissolved in 200 μL of nuclease-free water, and RNA was extracted immediately from the solution.

Paired primary tumor/adenoma tissue and nonmalignant adjacent mucosa were obtained from CRC and adenoma patients (at least 20 cm distant), collected during surgical resection and immediately immersed in RNAlater solution (Ambion). All tissues samples were stored at –80°C until use.

Extraction of Total RNA

Total RNA was extracted from all stool samples using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as previously described.[8,10] Total RNA from plasma EVs was extracted as described by Sabo et al[9] and Ferrero et al.[10] For tissue samples, total RNA was isolated using QIAzol (Qiagen) after tissue homogenization performed with ULTRA-TURRAX Homogenizer (IKA), followed by phenol/chloroform extraction according to the manufacturer's standard protocol.

Library Preparation for Small RNA Sequencing

Small RNA-seq libraries were prepared from RNA extracted from tissues, stool, and plasma EVs as previously described by Tarallo et al.[6] Briefly, the NEBNext Multiplex Small RNA Library Prep for Illumina (New England Biolabs, Inc) kit was used to convert small RNA transcripts into barcoded complementary DNA (cDNA) libraries. For each library, 6 μL of RNA (35 ng for EV RNA and 250 ng for tissue/stool RNA) was processed as the starting material. Each library was prepared with a unique indexed primer. Multiplex adapter ligations, RT primer hybridization, RT reaction, and PCR amplification were performed according to the manufacturer's protocol. After PCR amplification, the cDNA constructs were purified with the QIAQuick PCR Purification Kit (Qiagen), following the modifications suggested by the NEBNext Multiplex Small RNA Library Prep for Illumina protocol. Final libraries were loaded on the Bioanalyzer 2100 (Agilent Technologies) using the DNA High Sensitivity Kit (Agilent Technologies) according to the manufacturer's protocol. Libraries were pooled together (in 24-plex or 30-plex) and further purified with a gel size selection. A final Bioanalyzer 2100 run with the High Sensitivity DNA Kit (Agilent Technologies) allowed us to assess DNA library quality regarding size, purity, and concentration. The obtained libraries were subjected to the Illumina sequencing pipeline on Illumina HiSeq4000 and NextSeq500 sequencers (Illumina Inc).

Quantitative Real-Time Polymerase Chain Reaction

Five miRNAs of the final signature (miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p, and miR-1246) were validated with a different technique in 2 subsets of stool RNA from the IT cohort (n = 51), the CZ cohort (n = 45), and the FIT cohort (n 8) using the miRCURY LNA SYBR Green PCR kit (Qiagen) according to the manufacturer's instructions for plasma/serum. RT was performed using the miRCURY LNA RT kit (Qiagen) according to the manufacturer's instructions with the addition of 1 spike-in (UniSp6) to the RT reaction.

For qPCR, complement cDNA was diluted 1:30; 3 μL of 1:30 water-diluted cDNA products were mixed at 7 μL of miRCURY SYBR Green Mastermix and 1 μL of specific miRNA probe (Qiagen). All cDNA products were prepared in triplicate PCR reactions following the manufacturer's instructions. For quality control purposes, 1 RNA sample was measured twice, and a sample containing nuclease-free water and carrier RNA was profiled as the negative control. All the reactions were run on the ABI Prism 7900 Sequence Detection System (Applied Biosystems). A melt curve analysis was performed for the amplification specificity of each individual target per sample.

GenEx software (Multi-D) was used for data preprocessing, including interplate calibration, evaluation of isolation and RT efficiency, setting specific cutoffs for negative control miRNA cycle threshold (Ct) values, and triplicate averaging. The analyses were performed by calculating ΔCt values by global mean. The fold change was calculated as log2 – ΔΔCT between CRC and control samples. miRNAs with a Ct value of >38 were deemed to be not detected. To avoid biased inference due to qPCR nondetects (Ct value = 40), a left-censoring approach was used. Ct values of 40 were in fact substituted with the highest observed Ct value for a given miRNA.[11] Ct values were then normalized by subtracting the Ct value of the selected endogenous controls or the global mean Ct from each of the 5 miRNAs of interest. Differential miRNA expression was determined by logistic regression adjusted for age and smoking. The unadjusted *P* values of <.05 were considered as statistically significant because these analyses were hypothesis driven.

Bioinformatics and Statistical Analysis

Small RNA-seq pipeline analyses were performed using a previously published Docker-embedded software to guarantee the computational reproducibility of the analysis.[8] Trimmed reads were mapped against an in-house curated reference of human miRNAs based on miRbase v22 (Supplementary Table 1*A*). The alignment was performed using BWA algorithm v0.7.12.[12] miRNA levels were quantified using 2 methods called the "knowledge-based" and "position-based" methods, as described by Tarallo et al.[8] The sequences of the mature miRNAs were compared and, in the case of mature miRNAs characterized by identical sequences, the associated read counts were summed. An miRNA was considered as detected if supported by at least 20 normalized reads.

The age- and sex-adjusted differential expression analysis was performed using DESeq2 R package v1.22.2[13] using the likelihood ratio test method. For tissue samples, to test the significance of miRNA differential expression levels between CRC/adenoma tissue and matched adjacent nonmalignant colonic mucosa, a paired DESeq2 analysis was applied. An miRNA was considered differentially expressed (DEmiRNA) if associated with an adjusted *P* value of <.05 and a median number of reads of >20 in at least 1 study group. In each analysis in which the IT and CZ cohorts were analyzed together, the cohort variable was added to the DESeq2 model to adjust for the cohort batch effect.

Statistical analysis between continuous variables was performed using the Wilcoxon rank sum test or Kruskal-Wallis test. Statistical analysis between categorical variables was performed using the chi-square test.

Functional enrichment analysis was performed with RBiomirGS v0.2.12[14] in default settings and considering the validated miRNA-target interactions from miRTarBase and miRecord. A term was considered enriched if associated with an adjusted $P < .05$ and at least 2 target genes. The analysis was performed on the Kyoto Encyclopedia of Genes and Genomes (c2.cp.kegg.v7.5.1), Reactome (c2.cp.reactome.v7.5.1), WikiPathways (c2.cp.wikipathways.v7.5.1), Gene Ontology Biological Processes (c5.go.bp.v7.5.1), and Hallmark gene set libraries (h.all.v7.5.1) from MSigDB v7.5.1.[15] The analysis input was the average log2 fold change and combined adjusted *P* value computed by the differential expression analysis between the CRC and control groups of the IT cohort and CZ cohort.

Analysis of the copy number variation data from the COAD cohort of The Cancer Genome Atlas was performed by retrieving the GISTIC score from CBioPortal v4.1.15 (https://www.cbioportal.org/) considering the dataset named "Colorectal Adenocarcinoma (TCGA, PanCancer Atlas)."

Functional analysis of signature miRNA target genes was performed using Enrichr (version March 29th, 2021)[16] considering the validated targets provided by miRTarBase. A Gene Ontology Biological Process was considered enriched if associated with a $P < .001$. Because miR-607-5p was a novel miRNA identified in this study, its putative targets were predicted using miRanda v3.3a.[17] to scan the human 3′ untranslated region sequences from Ensembl v109. Among the 3807 potential targets identified, the top 100 genes characterized by the highest binding score were used for the analysis.

The correlation analysis between fecal miRNA levels and microbial abundances was performed by reanalyzing the small RNA-seq and shotgun metagenomic data from Thomas et al.[2] Preprocessing of metagenomic data was performed following the procedures described by Thomas et al[2] and Wirbel et al.[3] Specifically, raw reads quality controlled, adapter removal, and removal of human and PhiX reads were performed using the pipeline available at https://github.com/SegataLab/preprocessing. Then, taxonomic profiling was performed with MetaPhlAn3 in default settings with mpa_v30_CHOCOPhlAn_201901 as the markers database. Correlation analysis was performed using the Spearman method and graphically represented using the *corrplot* R package.

Explainable Machine Learning Approach

The 3-phase explainable ML approach to identify the minimal miRNA predictive signature is shown in Supplementary Figure 1. The 3 phases of the workflow were data preparation, feature selection and classification.

The data preparation phase has been designed to make the data usable to the ML approach and consists of (1) dataset loading and encoding, (2) dataset splitting in training and test sets, and finally (3) feature *z*-score normalization. The input data consist of a list of N individuals associated with the pathologic category, characterized by a set of covariates (eg, age and sex) and by a count matrix of dysregulated miRNAs. Once loaded and encoded, the dataset is represented by a matrix X paired with a vector Y. Matrix X is composed of N × M real numbers, where N is the number of individuals that are described by M features, which are either miRNAs or covariates. Vector Y is of length N as well and contains the encoded pathologic category of each participant represented in X.

The dataset is divided into training and test sets (with a given proportion of individuals, eg, 70% vs 30%). The former set is used to train ML models, and the latter is used only to evaluate the model performances. During the dataset split, a stratification of the participants according to the pathologic category and specific confounding covariates (eg, sex, age, disease stage) is performed. This guarantees that the proportion of pathologic categories of the whole dataset is maintained in both the training and test sets.

Finally, a *z*-score normalization is applied. The mean and standard deviation of all the features of the training set are estimated and used to normalize both the training and the test set.

The feature selection phase identifies the most relevant and nonredundant features in the distinction of the participants between groups of interest. To identify the *k*-best features from a given dataset, multiple selection criteria are available.[18] Specifically, filter methods assess feature relevance by computing a score between each feature and the target variable, whereas embedded strategies are based on learning algorithms that have built-in feature selection mechanisms. Hereby, the analysis of variance *F* test and mutual information were adopted as filter methods, whereas the embedded methods were based on logistic regression and random forest.

A repeated stratified *k*-fold cross-validation setting is adopted to apply the selection criteria on different subsamples of the training set to avoid—or at least reduce—data overfitting.

For this study, the whole procedure was repeated 30 times for any *k* from 1 to 25 to test feature sets composed of an increased number of DEmiRNAs. Each feature set was evaluated by a classification procedure, described later in this section, to identify its average performance.

The final selection was performed by means of a utility function, peak of the AUC(*k*), that guarantees the best balance between the AUC and the number of features selected—namely, to select the minimal number of miRNAs providing the best performance—that ultimately constitutes the miRNA predictive signature.

The classification phase is used to predict the qualitative response for a given individual to a category, according to the miRNA signature previously identified. Hereby, 3 classifiers were selected and applied independently: random forest,[19] logistic regression,[20] and gradient boosting.[21] The classifiers were applied with default values for the hyperparameters. Specifically, for the random forest classifier, the parameters were num_trees = 100 and criterion = entropy, penalty = l2 was selected for the logistic regression, and num_trees = 100 was set for the gradient boosting classifier. The set of patients to be classified was partitioned using a stratified 10-fold cross validation. For each classifier, 100 independent runs were performed. The performance metrics for each classifier—AUC, accuracy, precision, and recall—were computed as average metrics among all runs performed.

This approach was implemented in Python 3 using the following libraries: scikit-learn,[18] pandas, and matplotlib library[22] for ML algorithms, dataset representation, and data visualization, respectively.

Overview of the MicroRNA Content in the Analyzed Sample Types

Fecal samples from the Italian cohort and Czech cohort. From the analysis of small RNA-seq experiments, an average of 86.50% ± 10.03% of reads passed the preprocessing phase, and an average of 1.32% ± 2.22% of reads were aligned to human miRNAs. The observed percentage of aligned reads is in line with previous small RNA-seq analyses of fecal miRNA content.[6,8] Despite all miRNA annotations that were used for the differential expression analysis, a threshold of 20 normalized reads was used to define an miRNA as detected in a specific sample. Using this threshold, on average, 421.97 ± 222.07 (range, 86–1516) miRNAs were detected in each sample.

Fecal samples from the validation cohort. From the analysis of small RNA-seq experiments on the validation cohort, an average of 95.58% ± 2.88% of reads passed the preprocessing phase, and an average of 1.14% ± 1.34% of reads were aligned to human miRNAs. An average of 440.73 ± 217.94 (range, 75–1713) fecal miRNAs were detected in these samples.

Plasma extracellular vesicle samples. From the small RNA-seq experiments on plasma EV samples, an average of 91.41% ± 9.85% of sequencing reads passed the preprocessing phase and, on average, 20.12% ± 11.56% were assigned to human miRNA annotations. The average number of miRNAs detected in these samples was 309.69 ± 90.40 (range, 252–1213).

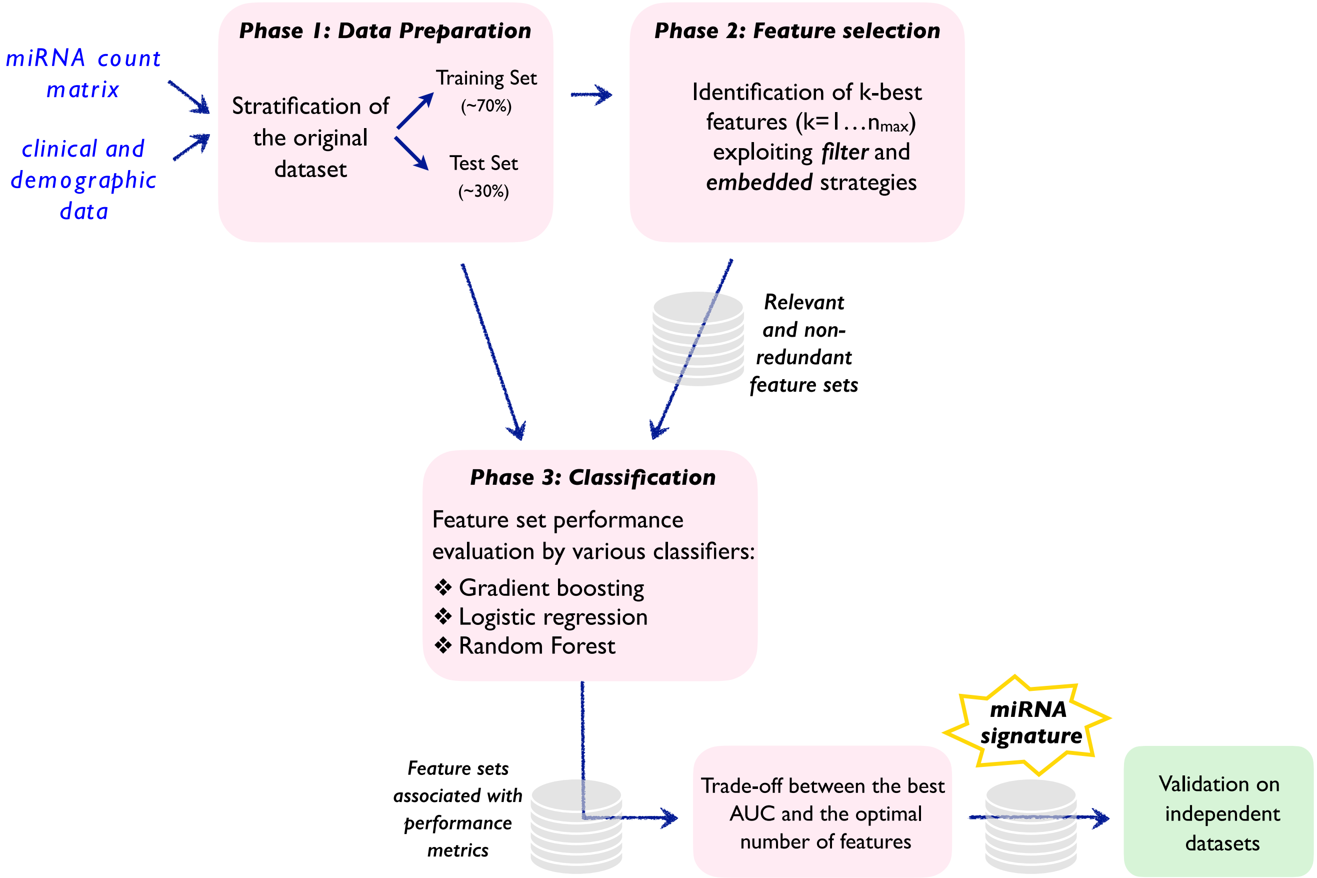
Tissue samples. In tissue samples, an average of 81.75% ± 13.01% sequencing reads were obtained from the preprocessing step, and among them, 68.56% ± 18.01% aligned on human miRNA annotations. On average, 581.84 ± 173.34 (range, 403–1997) miRNAs were detected in each sample.

Fecal immunochemical test leftover samples. From the small RNA-seq experiments on FIT leftover samples, an average of 90.30% ± 6.04% of sequencing reads passed the preprocessing phase, and, on average, 1.18% ± 0.49% were assigned to human miRNA annotations. The average number of miRNAs detected in these samples was 633.81 ± 41.07 (range, 541–744).

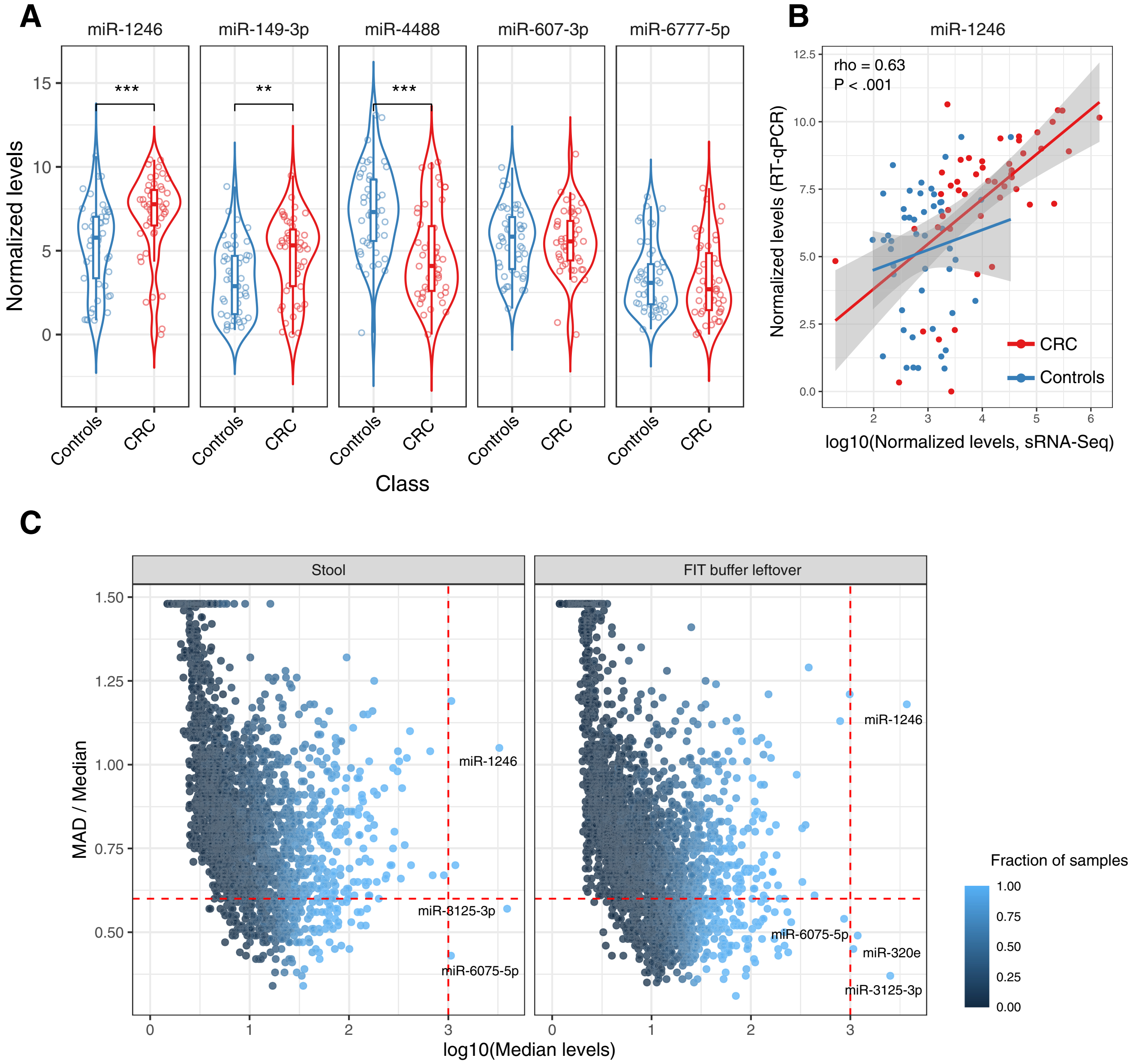
Supplementary References

1. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501–1504.
2. Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–678.
3. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–689.
4. Lin Y, Lau HC, Liu Y, et al. Altered mycobiota signatures and enriched pathogenic *Aspergillus rambellii* are associated with colorectal cancer based on multicohort fecal metagenomic analyses. Gastroenterology 2022;163:908–921.
5. Zwinsova B, Petrov VA, Hrivnakova M, et al. Colorectal tumour mucosa microbiome is enriched in oral pathogens and defines three subtypes that correlate with markers of tumour progression. Cancers (Basel) 2021;13(19):4799.
6. Tarallo S, Ferrero G, De Filippis F, et al. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. Gut 2022;71:1302–1314.
7. Francavilla A, Ferrero G, Pardini B, et al. Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk. Gut Microbes 2023;15(1):2172955.
8. Tarallo S, Ferrero G, Gallo G, et al. Altered fecal small RNA profiles in colorectal cancer reflect gut microbiome composition in stool samples. mSystems 2019;4(5):e00289-19.
9. Sabo AA, Birolo G, Naccarati A, et al. Small non-coding RNA profiling in plasma extracellular vesicles of bladder cancer patients by next-generation sequencing: expression levels of miR-126-3p and piR-5936 increase with higher histologic grades. Cancers (Basel) 2020;12(6):1507.
10. Ferrero G, Cordero F, Tarallo S, et al. Small non-coding RNA profiling in human biofluids and surrogate tissues from healthy individuals: description of the diverse and most represented species. Oncotarget 2018;9:3097–3111.
11. McCall MN, McMurray HR, Land H, et al. On non-detects in qPCR data. Bioinformatics 2014;30:2310–2316.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
14. Zhang J, Storey KB. RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. PeerJ 2018;6:e4262.
15. Liberzon A, Birger C, Thorvaldsdottir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–425.

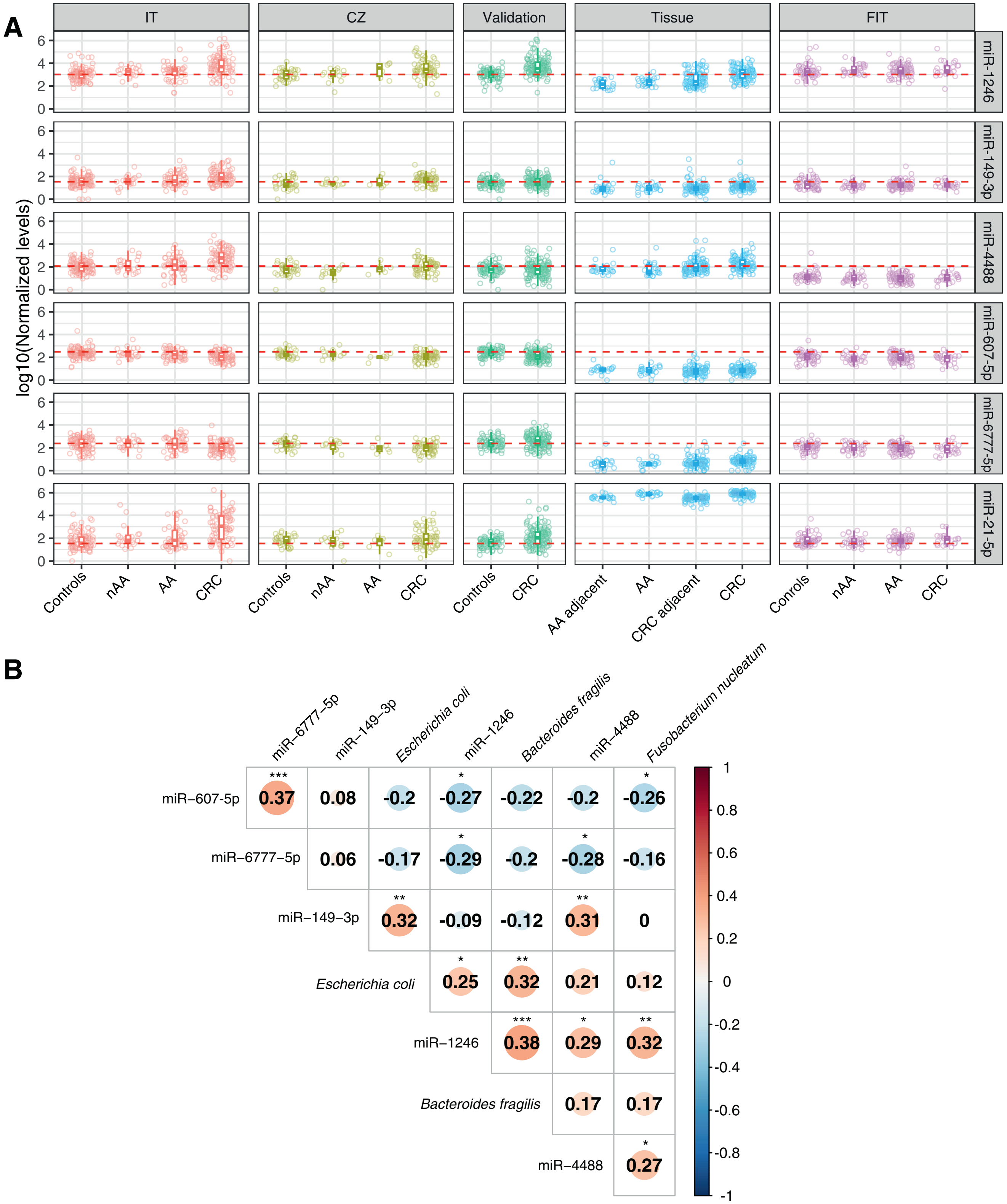
16. Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with Enrichr. Curr Protoc 2021;1(3):e90.
17. Betel D, Koppal A, Agius P, et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11(8):R90.
18. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learn Res 2011;12:2825–2830.
19. Breiman L. Random forests. Mach Learn 2001;45:5–32.
20. Fan RE, Chang KW, Hsieh CJ, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res 2008; 9:1871–1874.
21. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–1232.
22. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–95.


miRNA count matrix
clinical and demographic data
Phase 1: Data Preparation
Stratification of the original dataset
Training Set (~70%)
Test Set (~30%)
Phase 2: Feature selection
Identification of k-best features (k=1…n_max) exploiting *filter* and *embedded* strategies
Relevant and non-redundant feature sets
Phase 3: Classification
Feature set performance evaluation by various classifiers:
❖ Gradient boosting
❖ Logistic regression
❖ Random Forest
Feature sets associated with performance metrics
Trade-off between the best AUC and the optimal number of features
miRNA signature
Validation on independent datasets

Supplementary Figure 1. Schematic representation of the 3-phase explainable ML approach. An miRNA count matrix and the clinical/demographic data are the input data, and the best performing miRNA signature is the output.



Supplementary Figure 2. (*A*) Box plot showing the RT-qPCR normalized levels of the 5 miRNAs of the stool signature. *P* value by Wilcoxon rank sum test. ****P* < .001, ***P* < .01. (*B*) Scatterplot comparing the stool levels of miR-1246 measured by small RNA-seq (*x*-axis) and RT-qPCR (*y*-axis). The coefficient and significance of the Spearman correlation analysis is also reported. (*C*) Scatterplot reporting the median levels (*x*-axis) and the expression variability (as the ratio between median absolute deviation [MAD] and median, *y*-axis) of miRNAs measured in stool samples (*left plot*) or FIT buffer leftover (*right plot*) from the same subjects.



Supplementary Figure 3. (*A*) Box plots reporting, for each study cohort, the normalized levels of the 5 stool miRNAs belonging to our CRC-predictive signature. At the bottom, the levels of miR-21-5p are also reported. The red dashed lines refer to the median miRNA level measured in control individuals of the IT cohort. (*B*) Correlation plot representing the results of the Spearman correlation analysis between the levels of the 5 fecal miRNAs and *F nucleatum*, *E coli*, and *B fragilis* abundances by the reanalysis of data from Supplementary Reference.[2] The size of the dot is proportional to the absolute correlation coefficient. ****P* < .001; ***P* < .01; **P* < .05.

[***20***] **Budinská E**, Čarnogurská M, Ivković TC, Macháčková T, Boudná M, Pifková L, Slabý O, Bencsiková B, Popovici V. An invasion front gene expression signature for higher-risk patient selection in stage IIA MSS colon cancer. Front Oncol. 2024 Apr 19;14:1367231. doi: 10.3389/fonc.2024.1367231. PMID: 38706608; PMCID: PMC11066151.

TYPE Brief Research Report
PUBLISHED 19 April 2024
DOI 10.3389/fonc.2024.1367231

An invasion front gene expression signature for higher-risk patient selection in stage IIA MSS colon cancer

Eva Budinská[1], Martina Čarnogurská[1], Tina Catela Ivković[2], Táňa Macháčková[2,3], Marie Boudná[2,3], Lucie Pifková[2,3], Ondřej Slabý[2,3], Beatrix Bencsiková[4] and Vlad Popovici[1]*

[1]RECETOX, Faculty of Science, Masaryk University, Brno, Czechia, [2]Central European Institute of Technology, Masaryk University, Brno, Czechia, [3]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czechia, [4]Department of Comprehensive Cancer Care, Masaryk Memorial Cancer Institute, Brno, Czechia

OPEN ACCESS

EDITED BY
Sharon R. Pine,
University of Colorado Anschutz Medical Campus, United States

REVIEWED BY
Marahaini Musa,
University of Science Malaysia (USM), Malaysia
Marco Tonello,
Veneto Institute of Oncology (IRCCS), Italy
Chuanwen Fan,
Linköping University, Sweden
Sergio Facchini,
University of Campania Luigi Vanvitelli, Italy

*CORRESPONDENCE
Vlad Popovici
✉ vlad.popovici@recetox.muni.cz

RECEIVED 08 January 2024
ACCEPTED 18 March 2024
PUBLISHED 19 April 2024

CITATION
Budinská E, Čarnogurská M, Ivković TC, Macháčková T, Boudná M, Pifková L, Slabý O, Bencsiková B and Popovici V (2024) An invasion front gene expression signature for higher-risk patient selection in stage IIA MSS colon cancer.
Front. Oncol. 14:1367231.
doi: 10.3389/fonc.2024.1367231

Stage II colon cancer (CC) encompasses a heterogeneous group of patients with diverse survival experiences: 87% to 58% 5-year relative survival rates for stages IIA and IIC, respectively. While stage IIA patients are usually spared the adjuvant chemotherapy, some of them relapse and may benefit from it; thus, their timely identification is crucial. Current gene expression signatures did not specifically target this group nor did they find their place in clinical practice. Since processes at invasion front have also been linked to tumor progression, we hypothesize that aside from bulk tumor features, focusing on the invasion front may provide additional clues for this stratification. A retrospective matched case-control collection of 39 stage IIA microsatellite-stable (MSS) untreated CCs was analyzed to identify prognostic gene expression-based signatures. The endpoint was defined as relapse within 5 years vs. no relapse for at least 6 years. From the same tumors, three different classifiers (bulk tumor, invasion front, and constrained baseline on bulk tumor) were developed and their performance estimated. The baseline classifier, while the weakest, was validated in two independent data sets. The best performing signature was based on invasion front profiles [area under the receiver operating curve (AUC) = 0.931 (0.815–1.0)] and contained genes associated with KRAS pathway activation, apical junction complex, and heme metabolism. Its combination with bulk tumor classifier further improved the accuracy of the predictions.

KEYWORDS
colon cancer, invasion front, early stage, prognostic signature, stage II/MSS

1 Introduction

Despite important progress made in early detection and treatment over the last decades, colon cancer (CC) is still one of the major causes of death among all solid tumor cancers accounting for more than 600,000 deaths yearly (1). The TNM (tumor–node–metastasis) staging remains the cornerstone of patient management and outcome prediction, even though several other predictors have been proposed, including commercially available gene signatures, such as Oncotype Dx Colon (2), ColoPrint (3), and ColDX (4), or immune system scoring, such as Immunoscore Colon (5). Globally, stage II CC, accounting for 35%–40% of newly diagnosed cases (SEER Cancer Stat Facts: Colorectal Cancer; https://seer.cancer.gov/statfacts/html/colorect.html), has a good prognosis, with 5-year relative survival rates of 58%–87% (6). However, compared to other stages, it is more heterogeneous with low, intermediate, and high risk for metastatic dissemination subgroups, as recognized in the revised categorization (6). Microsatellite instability (MSI) or deficiency in DNA mismatch repair (dMMR) are characteristics of a low-risk group, with more than 90% 5-year overall survival (7). The high-risk (pT4bN0, stage IIC) or intermediate-risk microsatellite-stable (pT4aN0/MSS, stage IIB/MSS) patients are generally treated with adjuvant chemotherapy after curative surgical resection (8). The benefits from adjuvant therapy are not clear in these patients probably because direct evidence from clinical trials is still insufficient (9). However, the low-risk patients (pT3N0, stage IIA) are usually spared the adjuvant treatment, but still, approximately 13% of them will die within 5 years (6). Therefore, it is of utmost importance to develop better prognostic tools, eventually integrated with the TNM staging, targeting the earlier stage where the benefit from adjuvant treatment may potentially be significant.

All the transcriptomic signatures proposed so far considered whole-tumor sampling for RNA extraction. Still, mounting evidence suggests that processes taking place at the invasion front would be equally prognostic, if not even more. The activation of epithelial-to-mesenchymal transition (EMT) at aberrant expression of nuclear β-catenin as invasion front markers of tumor progression has been recognized previously (10, 11). Also, the infiltrative configuration of the invasion front and the presence of tumor budding have been recognized as additional prognostic parameters (12, 13). It has been proposed that the balance of pro- and anti-tumor factors at the invasion front may be decisive for tumor progression (14) and overexpression of *ZEB2* (an epithelial-to-mesenchymal transition-associated gene) as the invasion front has been identified as an independent prognostic factor in a general CC patient population (15). Additionally, the immune reaction scored along the invasion front could be used to stratify the CC patients into three distinct risk groups (5). In addition, the histopathologic characteristics of the reactive stroma at the invasion front have been shown to bear prognostic potential (16). Thus, it is of interest whether transcriptomics of the invasion front may bring novel discriminative markers that could improve patient stratification.

The goal of the present pilot study is twofold: to assess the prognostic utility of invasion front gene expression and develop a predictor of early relapse within the low-risk stage IIA/MSS colon cancers. From the same group of patients, we develop gene signatures from both bulk tumor (traditional tumor sampling) and tumor invasion front predicting the risk of relapse, and we compare their performance. As the study has a limited sample size, we opted for increasing the contrasts between the groups by selecting patients with relapse within 5 years vs. patients with no relapse for at least 6 years.

2 Materials and methods

2.1 Samples

This retrospective matched case-control study used tumor samples from patients with CC who underwent surgery at Masaryk Memorial Cancer Institute, Brno, Czech Republic, in the years 1998–2018. Inclusion criteria for this study were as follows: age >18 years, clinically and histopathologically confirmed diagnosis of primary CC, stage IIA (pathology T-stage 3, N0), microsatellite-stable primary tumors, and no adjuvant chemotherapy. Standard clinical and histopathological variables (TNM, grade, etc.) were retrieved for all patients. The "early relapse" group was defined as those patients experiencing a relapse within 5 years from the date of diagnosis, while the "no relapse" group consisted of patients who did not experience a relapse for at least 6 years. The relapse was defined as any disease recurrence or disease-related death except for any second primary cancers. To the extent possible, the two groups were further matched in terms of gender, age, and grade distribution. Failure of laboratory analyses (problematic sample preparation, low quality and/or quantity of isolated RNA, and low quality of expression data) was a reason for excluding these samples from the study.

From each tumor block, two different regions were sampled in adjacent sections: one representing the bulk tumor and one only the invasion front (Supplementary Figure 1). Each sample was profiled independently.

2.2 Expression profiling

The RNA extraction was performed from formalin-fixed paraffin-embedded histopathological slides using AllPrep DNA/RNA Kits (Qiagen, Hilden, Germany) according to manufacturer's instructions. The extracted RNA served as input for a GeneChip WT Pico Reagent Kit (Thermo Fisher Scientific, Waltham, MA, USA) for analysis of the transcriptome on whole-transcriptome arrays. Total RNA from HeLa cells provided in the kit was used as a positive control together with high-quality low-concentration RNA isolated from a serum as a low-input control. Clariom D Array for human samples (Thermo Fisher Scientific, Waltham, MA, USA) was used for target hybridization to capture both coding and multiple forms of non-coding RNA. Finally, the arrays were scanned using Affymetrix GeneChip Scanner 3000 7 G (Thermo Fisher Scientific, Waltham, MA, USA). All the samples complied with the quality control requirements, and none of the samples were excluded from the analysis.

2.3 Bioinformatics analyses

All resulting CEL files were processed using Bioconductor (17) (v.3.15) packages oligo (18) (v.1.60), affycoretools (v1.68), and, for Clariom D chip annotation, pd.clariom.d.human (v.3.14). For the quality control, we used AffyPLM (v.147) and imposed a maximal median Normalized Unscaled Standard Error (NUSE) of 1.12. All chips passing the quality control steps were normalized together using RMA (oligo) with core-probeset summarization. Further, the array data were summarized at gene level by selecting the most variable probeset per unique EntrezID, and entries corresponding to missing HUGO symbols, speculative transcripts, microRNA, and short non-coding RNA were discarded resulting in a reduced list of 28,663 unique genes.

For the identification of differentially expressed genes, we used linear models (limma package v.3.52.2) with a cut-off for false discovery rate (FDR) of 0.1. The pathways were scored in terms of enrichment in specific signatures using gene set enrichment analysis (GSEA) (19) as implemented in fgsea package (v.1.22.0). MSigDB (hallmark gene sets collection "H" v.7.4.1) (20) was used as the main source for gene sets and pathways. The gene expression classifiers were based on ElasticNet model as implemented in the R package caret (v.6.0). All data analyses were performed in R 4.3 (R Development Core Team, 2022).

The development of the predictive models required the following two major steps: feature generation and classifier training. These two steps were embedded in an external leave-one-out loop for estimating the performance. The main performance parameter of the model was the area under the receiver operating curve (AUC) with sensitivity and specificity also estimated and reported. For the feature generation step, we first selected the most predictive (in terms of AUC) and stable genes and grouped them into modules according to gene signatures from MSigDB (H-section). For estimating the stability of each gene, we generated b = 50 bootstraps of the current training set (at each iteration of the leave-one-out procedure) and recorded the AUC and direction of the association of the gene with the outcome. We defined the direction of a gene g as $d_g = +1$ if the average expression of the gene in the "early relapse" group was higher than in the "no relapse" group; otherwise, d_g = -1. The AUC for a gene was the average AUC from bootstrapping procedure, and the gene was considered stable if the direction of the association with the outcome was constant (over the b bootstraps). The gene modules were generated from MSigDB gene signatures by selecting the top five (in terms of AUC) subsets of n_g genes from each signature. The value of a module was defined as $\mathrm{n_g}^{-1}\ \Sigma\ d_g x^g$, where x_g is the expression of gene g in the module. By extension, the names of the gene modules were taken from the names of the corresponding signatures even though they no longer represented their de-/activation status. Then, an ElasticNet model was fitted on the top $\mathrm{n_f}$ gene modules. To minimize the chances of overfitting, the tested domain for n_g and n_f was limited to values 3, 4, and 5. No constraint was imposed on the number of times a gene could be selected in different modules (the signatures from MSigDB overlapped) nor on selecting only one module per gene signature. While this choice introduces potential redundancy in the model, it also improves its robustness.

To validate the modeling approach, we used two independent data sets (21) compatible with our experimental design (with the exception of unknown MSI status) publicly available from ArrayExpress under accession numbers E-MTAB-863 and E-MTAB-864, respectively. We further limited the set of genes to the intersection of the two platforms (Clariom D for our study and Affymetrix customized Almac array for the independent sets) resulting in 13,274 common symbols. Also, in the validation sets (denoted KEN1 and KEN2), we considered only the patients in our target group (pT3/pN0/pM0); the rest of the expression profiles were used for mitigating the differences between the two microarray platforms. The model built (and validated) on the restricted set of genes was considered as a baseline model. Additionally, as the two external data sets contained survival data as well, we estimated the probability of survival in the two predicted groups using the Kaplan–Meier estimator and tested for significant difference between the curves using the log-rank test.

The main analysis considered the full set of genes available on our platform (Clariom D) and concerned the two sampling regions as follows: bulk tumor and invasion front, respectively.

3 Results

In total, n = 39 patients were identified fitting the selection criteria [19 cases of early relapse (12 men) vs. 20 cases of no relapse (11 men)] resulting in 39 bulk tumor profiles. For the same patients, n = 35 [17 early relapse (11 men) and 18 no relapse (10 men)] good quality invasion front profiles were also generated. No statistically significant differences were found between groups regarding age, tumor location, or grade (Table 1).

3.1 Differentially expressed genes and pathways

The differential expression analyses of both bulk tumor and invasion front samples revealed no genes with significantly different expressions between early and no relapse groups after adjusting for multiple testing. Nevertheless, 204 and 333 genes had a significant (un-adjusted) p-value (≤ 0.01) within the bulk tumor and invasion front samples, respectively. Using the t-statistics estimated by limma as input for ordering the whole set of genes for GSEA, we identified a number of pathways/gene sets differentially activated between the early relapse and no relapse groups (Figure 1). The full list of significant (un-adjusted p-value) genes (p-value ≤ 0.01) is given in Supplementary Table 1 and the GSEA results in Supplementary Table 2.

TABLE 1 Basic patient population demographics for the training set.

	Early relapse (within 5 years)	No relapse (for at least 6 years)	p-Value	Test
N	19	20		
Age [mean (SD)]	69.5 (9.22)	68.9 (9.56)	0.849	Student's t-test
Gender				
Female Men	7 12	9 11	0.747	Fisher's exact test
Grade				
G2 G3	18 1	20 0	0.487	Fisher's exact test
Tumor site				
Right (including transverse colon) Left	14 5	12 8	0.501	Fisher's exact test

All patients were stage II/A, microsatellite stable.

3.2 Early relapse predictors

To validate the approach, we developed a baseline predictor of early relapse cases using a restricted set of genes common to the two platforms (Clariom D and Almac) and based on bulk tumor profiles. The optimal model used $n_f = 5$ gene modules each with $n_g = 4$ genes (see Table 2). Its estimated leave-one-out performance was $AUC_0 = 0.795(95\%CI = 0.625 - 0.964)$ (Figure 2A). The binary classification performance (for the default cut-off of 0.5) was sensitivity $Se = 0.737$ $(95\%CI = 0.488 - 0.908)$ and specificity $Sp = 0.8(95\%CI = 0563 - 0.943)$. At the same time, the observed performance on the validation sets was $AUC_{KEN1} = 0.731(95\%CI = 0.636 - 0.827)$ and $AUC_{KEN2} = 0.768(95\%CI = 0.612 - 0.874)$ being superior to the one reported elsewhere (21) (Supplementary Figure 2). The Kaplan–Meier curves for predicted groups ("no relapse" and "early relapse") were significantly different ($p < 0.001$) (Supplementary Figure 3).

For the genes in the modules, a positive sign (explicit or implicit) indicates its higher expression in the "early relapse" group, while the negative sign indicates the reverse situation.

With the modeling approach validated, we studied the predictive power of the profiles derived from bulk tumor and invasion front regions. First, we compared the univariate (per-gene) AUCs for bulk and invasion front profiles (Figure 2B, Supplementary Table 3) estimated using all samples. It was apparent that the invasion front expression profiles were more predictive with the top ranking genes having consistently higher univariate AUC (2%–5%). Also, there were almost twice as many genes from the invasion front with AUC > 0.7 than from bulk tumor profiles (Supplementary Table 3).

The predictors built from the bulk and invasion front profiles confirmed this tendency (Figure 2A): the leave-one-out estimated performance for invasion front was $AUC_i = 0.931(95\%CI = 0.815 - 1.0)(Se = 0.882, Sp = 0.833)$, superior to the bulk tumor performance: $AUC_b = 0.887(95\%CI = 0.750 - 1.0)(Se = 0.895, Sp = 0.75)$. The two models are given in Table 2 and further gene annotations in Supplementary Table 4.

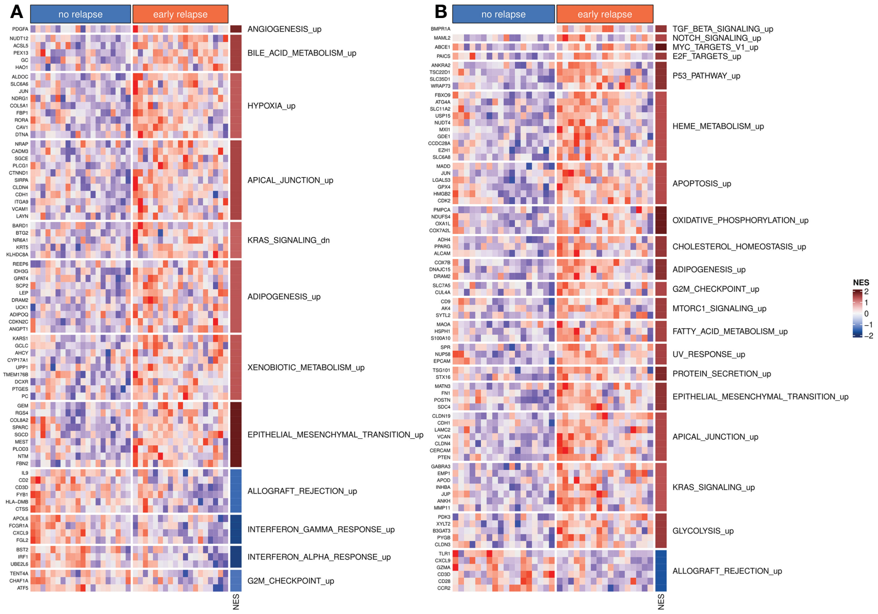


FIGURE 1
Differentially activated hallmark pathways. **(A)** Hallmark pathways and top differentially expressed genes from bulk tumor profiles. **(B)** Hallmark pathways and top differentially expressed genes from invasion front profiles. In both panels, *NES* indicates the normalized enrichment scores. The suffix "_up" or "_dn" indicates whether higher *NES* values correspond to set of gene sets that were activated ("_up") or inhibited ("_dn": down), respectively.

TABLE 2 Predictive models and their performance.

Model	Modules and coefficients	Module coefficient	Genes in modules	Leave-one-out performance estimates (with 95% confidence intervals)

Baseline model	INTERFERON_GAMMA_RESPONSE_up1
	INTERFERON_GAMMA_RESPONSE_up2
	INTERFERON_GAMMA_RESPONSE_up3
	TNFA_SIGNALING_VIA_NFKB_up1
	TNFA_SIGNALING_VIA_NFKB_up2
		1.0545
		−0.7575
		2.0305
		1.8225
		−0.9185
			LATS2 - IRF1 - TRIM14 - APOL6
			LATS2 - CXCL9 - TRIM14 - APOL6
			LATS2 - IRF1 - TRIM14 - CXCL9
			DUSP1 + LAMB3 - IRF1 - SLC2A6
			DUSP1 + JUN - IRF1 - SLC2A6
				AUC = 0.795 (0.625–0.964)
				Se = 0.737 (0.488–0.908)
				Sp = 0.8 (0.563–0.943)

Bulk tumor model	INFLAMMATORY_RESPONSE_up
	IL6_JAK_STAT3_SIGNALING_up
	APICAL_JUNCTION_up
		1.1161
		−0.1483
		0.6747
			EBI3 + KCNMB3 + TLR2 - IRF1 - TACR3
			EBI3 + HAX1 + TLR2 - IRF1 - CXCL9
			CLDN4 + LAYN + ITGA9 + NRAP + CADM3
				AUC = 0.887 (0.75–1.0)
				Se = 0.895 (0.669–0.987)
				Sp = 0.75 (0.509–0.913)

Invasion front model	APICAL_JUNCTION_up
	KRAS_SIGNALING_up
	HEME_METABOLISM_up
		0.1652
		0.1527
		0.0915
			VCAN + CLDN19 + PTEN + CDH1
			GABRA3 + APOD + JUP - TMEM100
			EZH1 + CCDC28A + FBXO9 + SLC6A8
				AUC = 0.931 (0.815–1.0)
				Se = 0.882 (0.636–0.985)
				Sp = 0.833 (0.586–0.964)

3.3 Combining predictors

We also compared the scores (posterior probabilities from ElasticNet models) produced by the two models (Figure 2C). The correlations (Pearson correlation: 0.564, Spearman correlation: 0.582) between the scores were modest, as was Cohen's kappa coefficient (κ = 0.484 between the class assignments based on these scores). This indicated a certain degree of complementarity between the two models, and we speculatively created an average score (from leave-one-out scores of matched tumor bulk and invasion front samples) and used it for predicting the groups. The new score indeed improved on all previous predictions—$AUC = 0.977(95\%CI = 0.907 - 1.0),Se = 0.941,Sp = 0.889$).

4 Discussion

The intermediate-risk group of patients with stage II colon cancer is heterogeneous in terms of survival experience: while most of the patients fare well without any adjuvant chemotherapy, others relapse much sooner. Reliably identifying the patients at risk for early relapse is, therefore, fundamental.

Our pilot study addressed two problems: First, developing a gene-based predictor for the stage IIA colon cancer patients who, despite being considered as low risk of relapse by current guidelines, are relapsing within 5 years. The second problem addressed aimed at investigating whether the invasion front is more predictive for the early relapse. Benefitting from a matched data set on which both bulk tumor and invasion front were profiled, we developed two predictive models. In our data, the invasion front model proved to be significantly superior to the bulk tumor model. This suggests that the dynamic changes happening on the contact border between the tumor and the normal tissue of the host may bear more information about the invasiveness potential of the tumor.

The targeted patient population appears to be rather homogeneous from the perspective of transcriptomics, with no gene significantly differentially expressed between "no relapse" and "early relapse" groups, after adjustment for multiple hypotheses testing. Nevertheless, several genes reached statistical significance when considered individually with more genes in the case of invasion front samples. Using the results from the differential expression analysis as input for gene set enrichment analyses, several significantly deregulated pathways/gene sets were identified. Some of them were common between bulk tumor and invasion front samples, most notably the epithelial-to-mesenchymal transition pathway, which was strongly up regulated in early relapse cases. Interestingly, the KRAS activation appeared in contrasting instances between the following two types of samples: in bulk tumors, the KRAS-down gene set was activated in the "early relapse" group, while in invasion front samples, the KRAS-up gene set was activated in the same group of patients, indicating a differential activation of KRAS between bulk tumor and invasion front regions within early CC.

The first predictor for early relapse established a baseline model and performance and validated the modeling approach. However, it was limited in the number of genes covered, as the two independent validation sets originated from an older microarray platform. Nevertheless, we were able to construct and validate a relatively strong classifier from bulk tumor profiles. The validation sets (21) were not selected for MSS, as this was not reported, but the baseline model performed close to the estimated performance. While the baseline classifier relied on five gene modules, the features selected by the algorithm referred to only two of the following MSigDB's pathways: interferon-gamma (INF-γ) and tumor necrosis factor-alpha (TNF-α) via nuclear factor-κβ, related to antitumor immunity and inflammatory processes, respectively. More interestingly, one gene—*IRF1* (interferon regulatory factor 1)—was common to both pathways (and to both bulk tumor models) and selected in four out of five modules being downregulated in the early relapse group. Upregulation of this gene was shown to be related to better survival and tumor radiosensitivity (22). We also note that the model could be further simplified to a model with only two modules (INF-γ and TNF-α) each of five genes; however, this combination was not foreseen when training the models (we imposed n_f = 3,4,V 5).

The same modeling approach was applied on tumor bulk and invasion front profiles considering all the genes present on our platform (still limited to the hallmark pathways of MSigDB). This led to the development of two models of which the invasion front

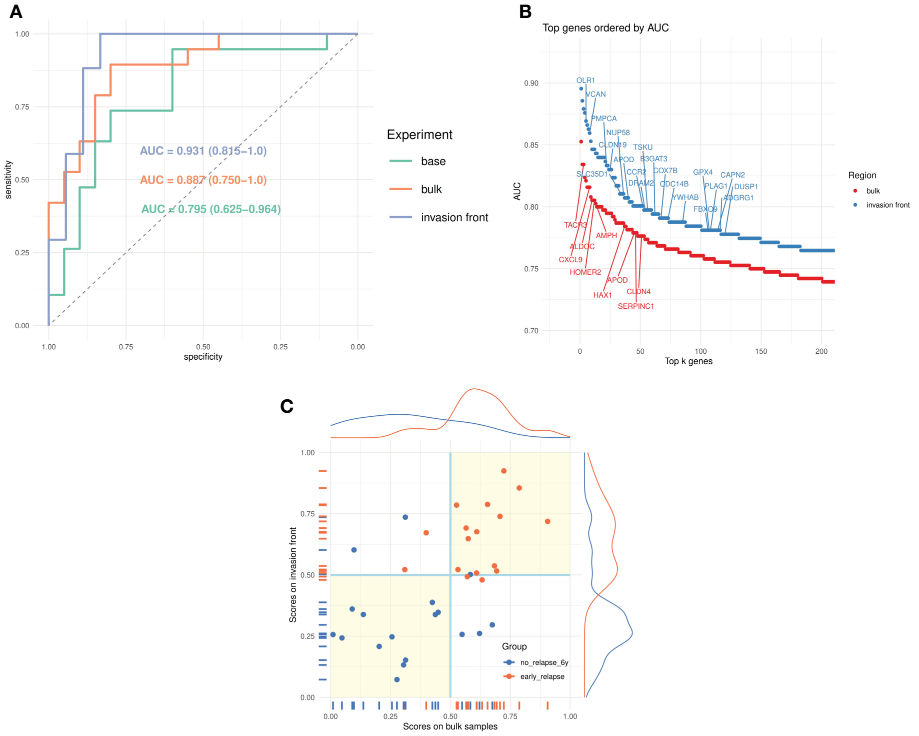


FIGURE 2
Prediction of early relapse. **(A)** Receiver operating characteristics (ROC) curves for the three models (baseline, bulk tumor, and invasion front) and the corresponding AUCs. **(B)** Univariate AUC, based on all samples, for top *k* = 200 genes from bulk tumor and invasion front expression profiles. The top genes (AUC > 0.775) from MSigDB hallmark signatures are marked. **(C)** Scatter plots of scores from bulk tumor and invasion front (35 samples) and their marginal distributions. The points are colored according to their true category, and the quadrants marked (light yellow background) indicate regions of agreement for the two classifiers.

signature had the best performance while both being superior to the baseline model. As the models were derived from tumor samples originating from the same patients, comparing the two allowed us to gain more insights into the predictive power of the invasion front. We first investigated the predictive utility (in terms of AUC) of each gene and found more genes from the invasion front having higher AUCs than from bulk tumors (see also Supplementary Table 3). While these results hinted toward more prognostic value of the invasion front signatures, it was the multivariable models (ElasticNets) that showed this being true in practice. Both models comprised of three gene modules with apical junction being a common term. However, the genes selected in the two "apical junction" modules were different with those from the invasion front pointing also toward EMT (*VCAN*) and estrogen receptor (*CDH1*). Also, we note the KRAS-related module present in the invasion front signature, which, corroborated with the results of GSEA (Figure 1; Supplementary Table 2), points toward a stronger KRAS pathway activation in early relapse patients. While specific mutations of the *KRAS* oncogene were shown to be predictive for overall survival in some studies (23, 24), they appeared not to be predictive for relapse-free survival (25). A more detailed annotation of all genes, with further references, is given in Supplementary Table 4. We also noted that the proposed marker gene for invasion front (15), *ZEB2*, was prognostic in our data as well, but with lower performance [$\text{AUC}_{\text{ZEB2}} = 0.716$ (0.521–0.910); Supplementary Table 3].

Our pilot study has some limitations as well: the invasion front signature could not be validated on external independent data because no similar data collections exist. We make our data publicly available to begin filling this gap. Second, the sample size did not allow for more analyses. For example, the observation that combining invasion front and bulk tumor signatures into a stronger predictor was made *post hoc*, and it would require another data set for its statistical assessment.

Another aspect pertains to the definition/delineation of the invasion front. We expect a relatively significant inter-observer variability. Thus, for the future results to be validated independently, a consensus must be reached between pathologists to stabilize the sampling regions.

In conclusion, our study proposes a novel invasion front-derived gene signature for predicting high-risk patients within the stage IIA colon cancer group. Its combination with bulk tumor signature further improved the prediction suggesting that a combined, dual sampling of core and border of the tumor may lead to a practical and precise predictor.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/arrayexpress/, E-MTAB-13695.

Ethics statement

The studies involving humans were approved by Research Ethics committee of Masaryk University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

EB: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft. MČ: Methodology, Writing – original draft. TI: Investigation, Methodology, Writing – original draft. TM: Data curation, Writing – original draft. MB: Data curation, Writing – original draft. LP: Data curation, Writing – original draft. OS: Conceptualization, Methodology, Writing – original draft. BB: Data curation, Funding acquisition, Investigation, Project administration, Writing – original draft. VP: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Supported by the Ministry of Health of the Czech Republic, grant nr. 19-03-00298. All rights reserved. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825410 (ONCOBIOME project). This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Acknowledgments

The authors thank the RECETOX Research Infrastructure (No LM2023069) financed by the Ministry of Education, Youth and Sports for supportive background. Supported by Ministry of Health of the Czech Republic, grant nr. 19-03-00298. All rights reserved. This work was supported from the European Union's Horizon 2020 research and innovation program under grant agreement No 857560.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2024.1367231/full#supplementary-material

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clin*. (2018) 68:394–424. doi: 10.3322/caac.21492

2. Gray RG, Quirke P, Handley K, Lopatin M, Magill L, Baehner FL, et al. Validation study of a quantitative multigene reverse transcriptase–polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *JCO*. (2011) 29:4611–9. doi: 10.1200/JCO.2010.32.8732

3. Kopetz S, Tabernero J, Rosenberg R, Jiang Z-Q, Moreno V, Bachleitner-Hofmann T, et al. Genomic classifier coloPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncol*. (2015) 20:127–33. doi: 10.1634/theoncologist.2014-0325

4. Niedzwiecki D, Frankel WL, Venook AP, Ye X, Friedman PN, Goldberg RM, et al. Association between results of a gene expression signature assay and recurrence-free interval in patients with stage II colon cancer in cancer and leukemia group B 9581 (Alliance). *JCO*. (2016) 34:3047–53. doi: 10.1200/JCO.2015.65.4699

5. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou F-S, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet*. (2018) 391:2128–39. doi: 10.1016/S0140-6736(18)30789-X

6. Gunderson LL, Jessup JM, Sargent DJ, Greene FL, Stewart AK. Revised TN categorization for colon cancer based on national survival outcomes data. *JCO*. (2010) 28:264–71. doi: 10.1200/JCO.2009.24.0952

7. Dienstmann R, Mason MJ, Sinicrope FA, Phipps AI, Tejpar S, Nesbakken A, et al. Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. *Ann Oncol*. (2017) 28:1023–31. doi: 10.1093/annonc/mdx052

8. Taieb J, Karoui M, Basile D. How I treat stage II colon cancer patients. *ESMO Open*. (2021) 6(4):100184. doi: 10.1016/j.esmoop.2021.100184

9. Argilés G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. (2020) 31:1291–305. doi: 10.1016/j.annonc.2020.06.022

10. Brabletz T, Hlubek F, Spaderna S, Schmalhofer O, Hiendlmeyer E, Jung A, et al. Invasion and metastasis in colorectal cancer: epithelial-mesenchymal transition, mesenchymal-epithelial transition, stem cells and β-catenin. *Cells Tissues Organs*. (2005) 179:56–65. doi: 10.1159/000084509

11. Bronsert P, Enderle-Ammour K, Bader M, Timme S, Kuehs M, Csanadi A, et al. Cancer cell invasion and EMT marker expression: a three-dimensional study of the human cancer-host interface: 3D cancer-host interface. *J Pathol*. (2014) 234:410–22. doi: 10.1002/path.4416

12. Compton CC, Fielding LP, Burgart LJ, Conley B, Cooper HS, Hamilton SR, et al. Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999. *Arch Pathol Lab Med*. (2000) 124:979–94. doi: 10.5858/2000-124-0979-PFICC

13. Graham RP, Vierkant RA, Tillmans LS, Wang AH, Laird PW, Weisenberger DJ, et al. Tumor budding in colorectal carcinoma: confirmation of prognostic significance and histologic cutoff in a population-based cohort. *Am J Surg Pathol*. (2015) 39:1340–6. doi: 10.1097/PAS.0000000000000504

14. Zlobec I, Lugli A. Invasive front of colorectal cancer: dynamic interface of pro-/anti-tumor factors. *World J Gastroenterol*. (2009) 15:5898–906. doi: 10.3748/wjg.15.5898

15. Kahlert C, Lahes S, Radhakrishnan P, Dutta S, Mogler C, Herpel E, et al. Overexpression of ZEB2 at the invasion front of colorectal cancer is an independent prognostic marker and regulates tumor invasion *in vitro*. *Clin Cancer Res*. (2011) 17:7654–63. doi: 10.1158/1078-0432.CCR-10-2816

16. Martin B, Grosser B, Kempkens L, Miller S, Bauer S, Dhillon C, et al. Stroma AReactive invasion front areas (SARIFA)-A new easily to determine biomarker in colon cancer-results of a retrospective study. *Cancers*. (2021) 13(19):4880. doi: 10.4324/9781003101857

17. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. (2015) 12:115–21. doi: 10.1038/nmeth.3252

18. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. (2010) 26:2363–7. doi: 10.1093/bioinformatics/btq431

19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102

20. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst*. (2015) 1:417–25. doi: 10.1016/j.cels.2015.12.004

21. Kennedy RD, Bylesjo M, Kerr P, Davison T, Black JM, Kay EW, et al. Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue. *JCO*. (2011) 29:4620–6. doi: 10.1200/jco.2011.35.4498

22. Xu X, Wu Y, Yi K, Hu Y, Ding W, Xing C. IRF1 regulates the progression of colorectal cancer via interferon–induced proteins. *Int J Mol Med*. (2021) 47:104. doi: 10.3892/ijmm.2021.4937

23. Imamura Y, Morikawa T, Liao X, Lochhead P, Kuchiba A, Yamauchi M, et al. Specific mutations in KRAS codons 12 and 13, and patient prognosis in 1075 BRAF wild-type colorectal cancers. *Clin Cancer Res*. (2012) 18:4753–63. doi: 10.1158/1078-0432.CCR-11-3210

24. Jones RP, Sutton PA, Evans JP, Clifford R, McAvoy A, Lewis J, et al. Specific mutations in KRAS codon 12 are associated with worse overall survival in patients with advanced and recurrent colorectal cancer. *Br J Cancer*. (2017) 116:923–9. doi: 10.1038/bjc.2017.37

25. Popovici V, Budinska E, Bosman FT, Tejpar S, Roth AD, Delorenzi M. Context-dependent interpretation of the prognostic value of BRAF and KRAS mutations in colorectal cancer. *BMC Cancer*. (2013) 13:439. doi: 10.1186/1471-2407-13-439