Molecular portraits of colorectal cancer morphological regions

Eva Budinská: eva.budinska@recetox.muni.cz, RECETOX, Faculty of Science, Masarykova Univerzita, Kotlarska 2, 602 00 Brno, Czech Republic

Martina Hrivňáková: martina.hrivnakova@recetox.muni.cz, RECETOX, Faculty of Science, Masarykova Univerzita, Kotlarska 2, 602 00 Brno, Czech Republic

Tina Catela Ivkovic: tina.catela_ivkovic@med.lu.se, Central European Institute of Technology, Masarykova Univerzita, 625 00 Brno, Czech Republic

Marie Madrzyk: marie.madrzyk@ceitec.muni.cz, Central European Institute of Technology, Masarykova Univerzita, 625 00 Brno, Czech Republic

Rudolf Nenutil: nenutil@mou.cz, Masaryk Memorial Cancer Institute, 656 53 Brno, Czech Republic

Beatrix Bencsiková: bencsikova@mou.cz, Masaryk Memorial Cancer Institute, 656 53 Brno, Czech Republic

Dagmar Al Tukmachi: dagmar.altukmachi@ceitec.muni.cz, Central European Institute of Technology, Masarykova Univerzita, 625 00 Brno, Czech Republic

Michaela Ručková: michaela.ruckova@ceitec.muni.cz, Central European Institute of Technology, Masarykova Univerzita, 625 00 Brno, Czech Republic

Lenka Zdražilová Dubská: ldz@mail.muni.cz, Faculty of Medicine, Masarykova Univerzita, 625 00 Brno, Czech Republic

Ondřej Slabý: ondrej.slaby@oslaby.med.muni.cz, Central European Institute of Technology, Department of Biology, Faculty of Medicine, Masarykova Univerzita, 625 00 Brno, Czech Republic

Josef Feit: jfeit@ics.muni.cz, Department of Pharmacology and Toxicology, Faculty of Pharmacy, Masarykova Univerzita, 625 00 Brno, Czech Republic

Mihnea-Paul Dragomir: mihnea.dragomir@charite.de, Institute of Pathology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, 10117 Germany; Berlin Institute of Health, Berlin, Germany; German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany

Petra Borilova Linhartova: petra.linhartova@recetox.muni.cz, RECETOX, Faculty of Science, Masarykova Univerzita, Kotlarska 2, 602 00 Brno, Czech Republic

Sabine Tejpar: sabine.tejpar@kuleuven.be, Faculty of Medicine, Digestive Oncology Unit, Katholieke Universiteit Leuven, 3000 Leuven, Belgium

Vlad Popovici: vlad.popovici@recetox.muni.cz, RECETOX, Faculty of Science, Masarykova Univerzita, Kotlarska 2, 602 00 Brno, Czech Republic

Running title: Molecular portraits of morphotypes

Keywords: colorectal cancer, tumor morphology, intra-tumor heterogeneity, transcriptomics

Additional information:

Funding: The authors acknowledge funding from the Czech Science Foundation (GAČR), grant no. GA19-08646S (EB, MH, VP).

Corresponding author:

Vlad Popovici

Fac. of Science – RECETOX
Kamenice 34
625 00 Brno
Tel: +420 549 49 6848
Email: vlad.popovici@recetox.muni.cz

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.

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 (1), 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 (2–7) with consensus molecular subtypes (CMS)(8) 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 (9). Other studies combined genomics and transcriptomics data and an alternative classification emerged (10).

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 (11,12). 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 (13) 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, Budinská *et al.* (3) 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 (14,15). Müller *et al.* (16) 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 (17). 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 (18) and may partially explain the low proportion of biomarkers that reach clinical relevance (19).

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 (3), 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 (Supp. Fig. SF1), each case being represented by one or several regions. 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 Supplemental Table ST1. The only statistically significant associations found were between MU or TB and grade 3 tumors, and SE and lower grade tumors (p=0.019, Supplemental Table ST2), 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, Supplemental tables ST3-4).

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" (20) 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 (21) and colon differentiated epithelial cells (22) 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 (21) 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 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 (23)) and cancer associated fibroblasts (CAFs) (24,25) (see Supplemental table ST3 and ST4 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 (25) – 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 (23), we observed even finer differences between morphotypes within all three cell type populations and their programs (Supp. Fig. SF2 A-C). 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 Supplemental Table ST4.

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 (Supplemental Table ST5 contains all DEGs). This allowed the identification of all molecular programs significantly de-/activated in each

morphotype (Table 1, Figure 3, Supplemental Table ST6). 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, Supplemental Figure SF3). 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 Supplemental Tables ST5-6). 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) (Supplemental Tables ST7-8).

Several macro dissected 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 (Supplemental Tables ST9-10). 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 Supplemental Figure SF4).

We also performed comparison between all pairs of morphotypes (Supplemental Tables ST11-12). 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 (ST11) and activated molecular programs (ST12) 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 (3), 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 macrodissected morphological regions. We predicted both the consensus (CMS)(8) and intrinsic (iCMS) (9) 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, Supplemental Figure SF7). More importantly, this heterogeneity was also observed intra-tumoral, with regions within the same tumor section being assigned to different subtypes (Figures 4A, 5, Supplemental Figure SF5A-B).

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, Supplemental Figure SF7) 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%).

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 Supplemental Figure SF5A-B).

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 (Supplemental Figure SF6B). The differences were observed even within tumor (Figure 5), with some of the supposedly sensitive tumors (whole tumor scoring) having regions of predicted resistance (Supplemental Figure SF6A-B).

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 (3). 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 (3). 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, (3)), 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 (23,25) 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 (26), 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 (27). Further, the poor prognostic associated with CMS4 could be explained by the stromal component: both (28) and (29) 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 (3), 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" (30). 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 $\leq$ 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 (31). 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 (32), *FLRT3* – a proapoptotic gene which, when overexpressed, inhibits EMT (33), *SETSIP* – gene coding chromatin-binding protein capable of participating in fibroblast reprogramming and differentiation into epithelial cells (34), *E2F7* – up-regulated by *p53* in response to DNA damage (35), *CXCL14* (downregulated) - depending on the cell of origin can have both tumor suppressive or supporting role (36), *SEMA5A* (downregulated) gene – proposed as prognostic marker in CRC (37). 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 (38), and *MUC2* (intestinal mucin) whose downregulation has been suggested a marker of adverse outcome (39). 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 (40). Still in DE, we found as top overexpressed genes *PIEZO2* – a paralog of *PIEZO1* which is involved in colorectal cancer metastases (41), *SLIT3* – a member of the Slit/Robo pathway, a major regulator of several oncogenic pathways and potential therapeutic target (42), and *OLFML2B* – a potential biomarker for resistance to MEK inhibitors (43). SE morphotype had only one specific gene overexpressed at FDR $\leq$ 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 (44), *GLIPR2* which participates in positive regulation of ERK1/2 cascade and EMT transition (45), or the *HOXA9A* gene, the overexpression of which was suggested to contribute to stem cell overpopulation responsible for development of CRC (46) 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 anti-cancer agents (5-FU and bevacizumab) (47). 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 costs 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, i.e. whether sufficiently discriminatory features could be extracted for estimating the proportions of the morphotypes from whole tumor profiles.

Materials and Methods

Sample preparation

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 20x 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. The study was approved by the ethical committee of Masaryk Memorial Cancer Institute, number 2018/861/MOU.

RNA samples were hybridized on Clariom™ D (human) microarrays (Thermo Fisher Scientific, Waltham, MA, USA) and data preprocessed using standard Bioconductor (RRID:SCR_006442) packages (see Supplemental Methods).

The data generated in this study are publicly available in ArrayExpress under accession number E-MTAB-12599.

Bioinformatics analyses

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) (48) as implemented in `fgsea` package (v.1.22.0). For scoring the signatures in individual samples, we used gene score variation analysis (GSVA) (49) implemented in `GSVA` package (v.1.44.1). `MSigDB` (RRID:SCR_016863) (all collections: H, C1-8; v.7.4.1) (50) 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) (20,23–25) (ii) epithelial cells (21–23), and (iii) immune cells (20,23) were used (see Supplemental Table 3 for full list). The consensus molecular subtypes were predicted using `CMSCaller` (51) (v.2.0.1) and the intrinsic epithelial subtypes (9) using the signatures therein (P. Tsantoulis, personal communication, July 2022). The cellular mixture of various tumoral regions was explored computationally using `quanTIseq` (52) (for immune cells) and `ESTIMATE` (53) (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 Supplemental Table ST13) 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. More detailed description of methods can be found in Supplemental material.

Acknowledgements

We would like to 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 (GAČR) 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.

Competing interests

The authors declare no potential conflicts of interest.

References

1. Amin MB, American Joint Committee on Cancer, American Cancer Society, editors. AJCC cancer staging manual. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP ; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR-Technical editor ; Laura R. Meyer, CAPM-Managing editor. Chicago IL: American Joint Committee on Cancer, Springer; 2017. 1024 p.

2. Perez Villamil B, Romera Lopez A, Hernandez Prieto S, Lopez Campos G, Calles A, Lopez Asenjo JA, et al. Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. BMC Cancer. 2012 Jun 19;12(1):260.

3. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer: Gene expression heterogeneity in colorectal cancer. J Pathol. 2013 Sep;231(1):63–76.

4. 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.

5. 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.

6. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med. 2013 May;19(5):619–25.

7. 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.

8. 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.

9. Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat Genet. 2022 Jul;54(7):963–75.

10. 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.

11. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell & Bioscience. 2019 Jun 26;9(1):53.

12. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial transcriptomics. Nature. 2021 Aug;596(7871):211–20.

13. Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology. 2007 Jan;50(1):113–30.

14. Budinska E, Bosman F, Popovici V. Experiments in molecular subtype recognition based on histopathology images. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) [Internet]. Prague, Czech Republic: IEEE; 2016 [cited 2020 Apr 21]. p. 1168–72. Available from: http://ieeexplore.ieee.org/document/7493474/

15. 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):2002–9.

16. Müller MF, Ibrahim AEK, Arends MJ. Molecular pathological classification of colorectal cancer. Virchows Arch. 2016 Aug;469(2):125–34.

17. Bosman FT, World Health Organization, International Agency for Research on Cancer, editors. WHO classification of tumours of the digestive system. 4th ed. Lyon: International Agency for Research on Cancer; 2010. 417 p. (World Health Organization classification of tumours).

18. Dunne PD, McArt DG, Bradley CA, O'Reilly PG, Barrett HL, Cummins R, et al. Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. Clin Cancer Res. 2016 Aug 15;budinska(16):4095–104.

19. Stewart JP, Richman S, Maughan T, Lawler M, Dunne PD, Salto-Tellez M. Standardising RNA profiling based biomarker application in cancer—The need for robust control of technical variables. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2017 Aug;1868(1):258–72.

20. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, et al. Stromal contribution to the colorectal cancer transcriptome. Nat Genet. 2015 Apr;47(4):312–9.

21. Kosinski C, Li VSW, Chan ASY, Zhang J, Ho C, Tsui WY, et al. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proceedings of the National Academy of Sciences. 2007 Sep 25;104(39):15418–23.

22. Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, Rossell D, et al. The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse. Cell Stem Cell. 2011 May;8(5):511–24.

23. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, et al. Spatially organized multicellular immune hubs in human colorectal cancer. Cell. 2021 Sep 2;184(18):4734-4752.e20.

24. Khaliq AM, Erdogan C, Kurt Z, Turgut SS, Grunvald MW, Rand T, et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biology. 2022 May 11;23(1):113.

25. Kieffer Y, Hocine HR, Gentric G, Pelon F, Bernard C, Bourachot B, et al. Single-Cell Analysis Reveals Fibroblast Clusters Linked to Immunotherapy Resistance in Cancer. Cancer Discovery. 2020 Sep 1;10(9):1330–51.

26. Desbois M, Wang Y. Cancer-associated fibroblasts: Key players in shaping the tumor immune microenvironment. Immunological Reviews. 2021;302(1):241–58.

27. Loughrey MB, Fisher NC, McCooey AJ, Dunne PD. Comment on "Identification of EMT-related high-risk stage II colorectal cancer and characterisation of metastasis-related genes" | British Journal of Cancer [Internet]. [cited 2022 Feb 5]. Available from: https://www.nature.com/articles/s41416-020-01213-9

28. Roseweir AK, Park JH, Hoorn S ten, Powell AG, Aherne S, Roxburgh CS, et al. Histological phenotypic subtypes predict recurrence risk and response to adjuvant chemotherapy in patients with stage III colorectal cancer. The Journal of Pathology: Clinical Research. 2020;6(4):283–96.

29. ten Hoorn S, de Back TR, Sommeijer DW, Vermeulen L. Clinical Value of Consensus Molecular Subtypes in Colorectal Cancer: A Systematic Review and Meta-Analysis. JNCI: Journal of the National Cancer Institute. 2022 Apr 1;114(4):503–16.

30. De Palma FDE, D'Argenio V, Pol J, Kroemer G, Maiuri MC, Salvatore F. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. Cancers. 2019 Jul;11(7):1017.

31. Zhang Q, Zhang C, Ma JX, Ren H, Sun Y, Xu JZ. Circular RNA PIP5K1A promotes colon cancer development through inhibiting miR-1273a. World Journal of Gastroenterology. 2019 Sep 21;25(35):5300–9.

32. Liu P, Wang X, Pan L, Han B, He Z. Prognostic Significance and Immunological Role of FBXO5 in Human Cancers: A Systematic Pan-Cancer Analysis. Frontiers in Immunology [Internet]. 2022 [cited 2022 Dec 21];13. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2022.901784

33. Yang M, Li D, Jiang Z, Li C, Ji S, Sun J, et al. TGF-β-Induced FLRT3 Attenuation Is Essential for Cancer-Associated Fibroblast–Mediated Epithelial–Mesenchymal Transition in Colorectal Cancer. Molecular Cancer Research. 2022 Aug 5;20(8):1247–59.

34. Margariti A, Winkler B, Karamariti E, Zampetaki A, Tsai T neng, Baban D, et al. Direct reprogramming of fibroblasts into endothelial cells capable of angiogenesis and reendothelialization in tissue-engineered vessels. Proceedings of the National Academy of Sciences. 2012 Aug 21;109(34):13793–8.

35. Carvajal LA, Hamard PJ, Tonnessen C, Manfredi JJ. E2F7, a novel target, is up-regulated by p53 and mediates DNA damage-dependent transcriptional repression. Genes Dev. 2012 Jul 15;26(14):1533–45.

36. Westrich JA, Vermeer DW, Colbert PL, Spanos WC, Pyeon D. The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. Molecular Carcinogenesis. 2020;59(7):794–806.

37. Demirkol S, Gomceli I, Isbilen M, Dayanc BE, Tez M, Bostanci EB, et al. A Combined ULBP2 and SEMA5A Expression Signature as a Prognostic and Predictive Biomarker for Colon Cancer. Journal of Cancer. 2017 Apr 9;8(7):1113–22.

38. Patel A, Tripathi G, McTernan P, Gopalakrishnan K, Ali O, Spector E, et al. Fibroblast growth factor 7 signalling is disrupted in colorectal cancer and is a potential marker of field cancerisation. Journal of Gastrointestinal Oncology [Internet]. 2019 Jun [cited 2022 Dec 21];10(3). Available from: https://jgo.amegroups.com/article/view/27514

39. Betge J, Schneider NI, Harbaum L, Pollheimer MJ, Lindtner RA, Kornprat P, et al. MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: expression profiles and clinical significance. Virchows Arch. 2016 Sep 1;469(3):255–65.

40. Ueno H, Ishiguro M, Nakatani E, Ishikawa T, Uetake H, Murotani K, et al. Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial). British Journal of Cancer. 2021 Mar;124(6):1088–97.

41. Sun Y, Li M, Liu G, Zhang X, Zhi L, Zhao J, et al. The function of Piezo1 in colon cancer metastasis and its potential regulatory mechanism. J Cancer Res Clin Oncol. 2020 May 1;146(5):1139–52.

42. Gara RK, Kumari S, Ganju A, Yallapu MM, Jaggi M, Chauhan SC. Slit/Robo pathway: a promising therapeutic target for cancer. Drug Discovery Today. 2015 Jan 1;20(1):156–64.

43. Hu P, Zhang X, Li Y, Xu L, Qiu H. Pan-Cancer Analysis of OLFML2B Expression and Its Association With Prognosis and Immune Infiltration. Frontiers in Genetics [Internet]. 2022 [cited 2022 Dec 21];13. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2022.882794

44. Lee AM, Shi Q, Pavey E, Alberts SR, Sargent DJ, Sinicrope FA, et al. DPYD Variants as Predictors of 5-fluorouracil Toxicity in Adjuvant Colon Cancer Treatment (NCCTG N0147). JNCI: Journal of the National Cancer Institute. 2014 Dec 1;106(12):dju298.

45. Kang HN, Oh SC, Kim JS, Yoo YA. Abrogation of Gli3 expression suppresses the growth of colon cancer cells via activation of p53. Experimental Cell Research. 2012 Mar 10;318(5):539–49.

46. Osmond B, Facey COB, Zhang C, Boman BM. HOXA9 Overexpression Contributes to Stem Cell Overpopulation That Drives Development and Growth of Colorectal Cancer. International Journal of Molecular Sciences. 2022 Jan;23(12):6799.

47. Kang J, D'Andrea AD, Kozono D. A DNA Repair Pathway–Focused Score for Prediction of Outcomes in Ovarian Cancer Treated With Platinum-Based Chemotherapy. JNCI: Journal of the National Cancer Institute. 2012 May 2;104(9):670–81.

48. 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. 2005 Oct 25;102(43):15545–50.

49. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013 Jan 16;14(1):7.

50. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015 Dec 23;1(6):417–25.

51. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017 Nov 30;7(1):16618.

52. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Medicine. 2019 May 24;11(1):34.

53. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013 Dec;4(1):2612.

Tables

Table 1. Results of comparison of each morphotype (and the two non-tumoral regions) with the average profile. The table shows 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 Supplemental Tables ST5 and ST6.

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

Figures

Figure 1

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).

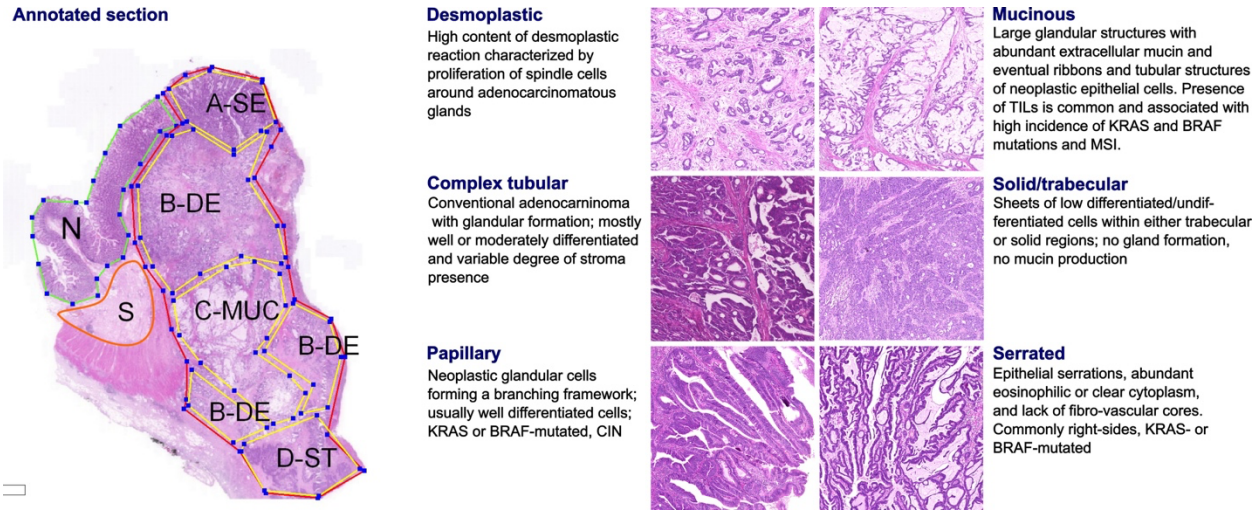


Figure 2

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 (24,25).

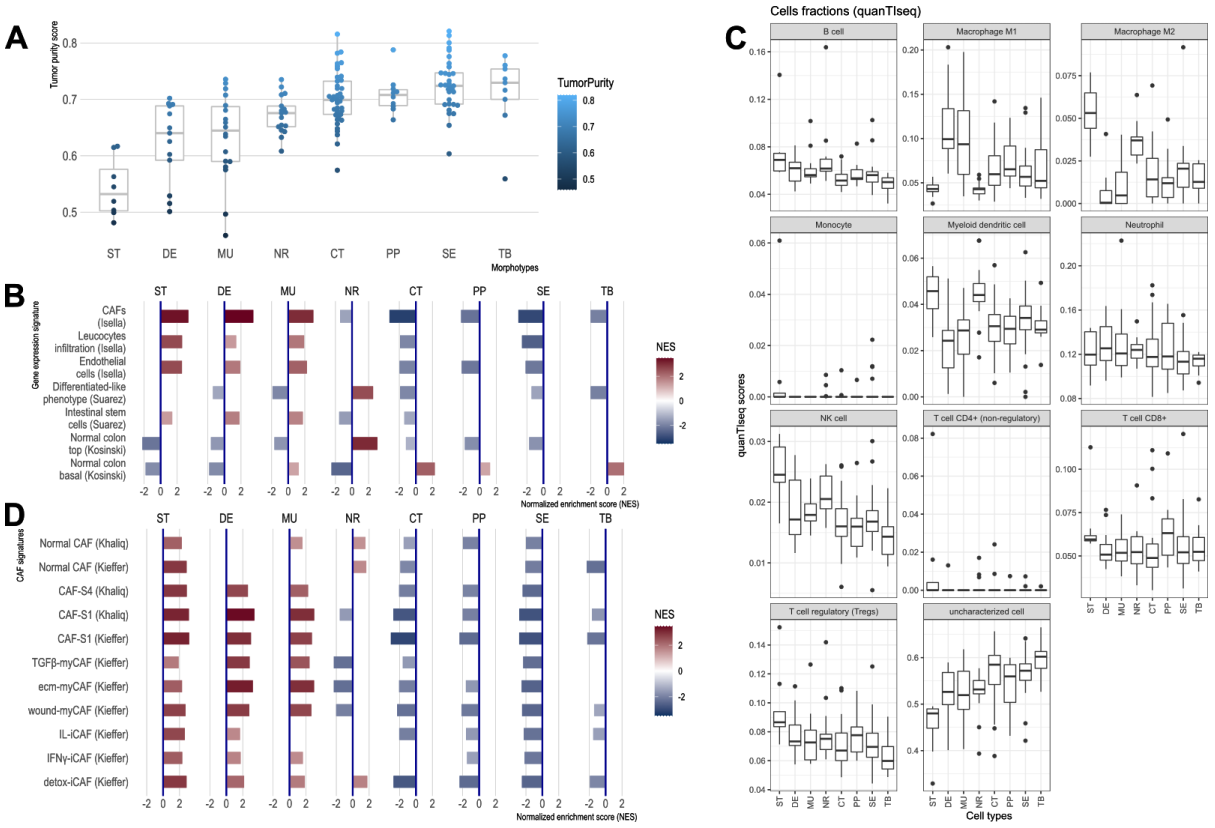


Figure 3

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 Supplemental Figure SF3. (C) Heatmap of top 5 up- and down-regulated genes for each of the six morphotypes.

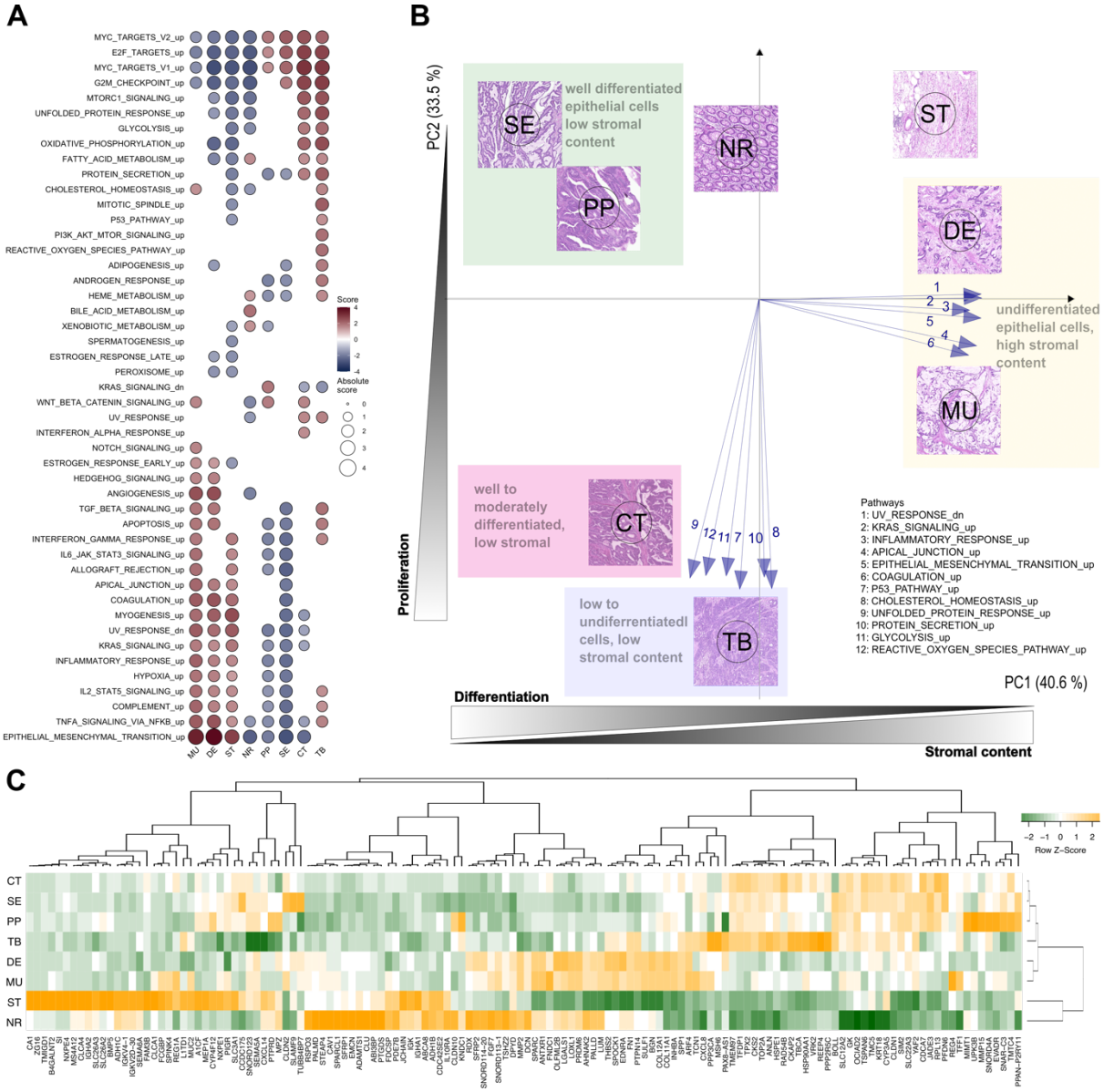


Figure 4

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 Supp.Fig. 7.

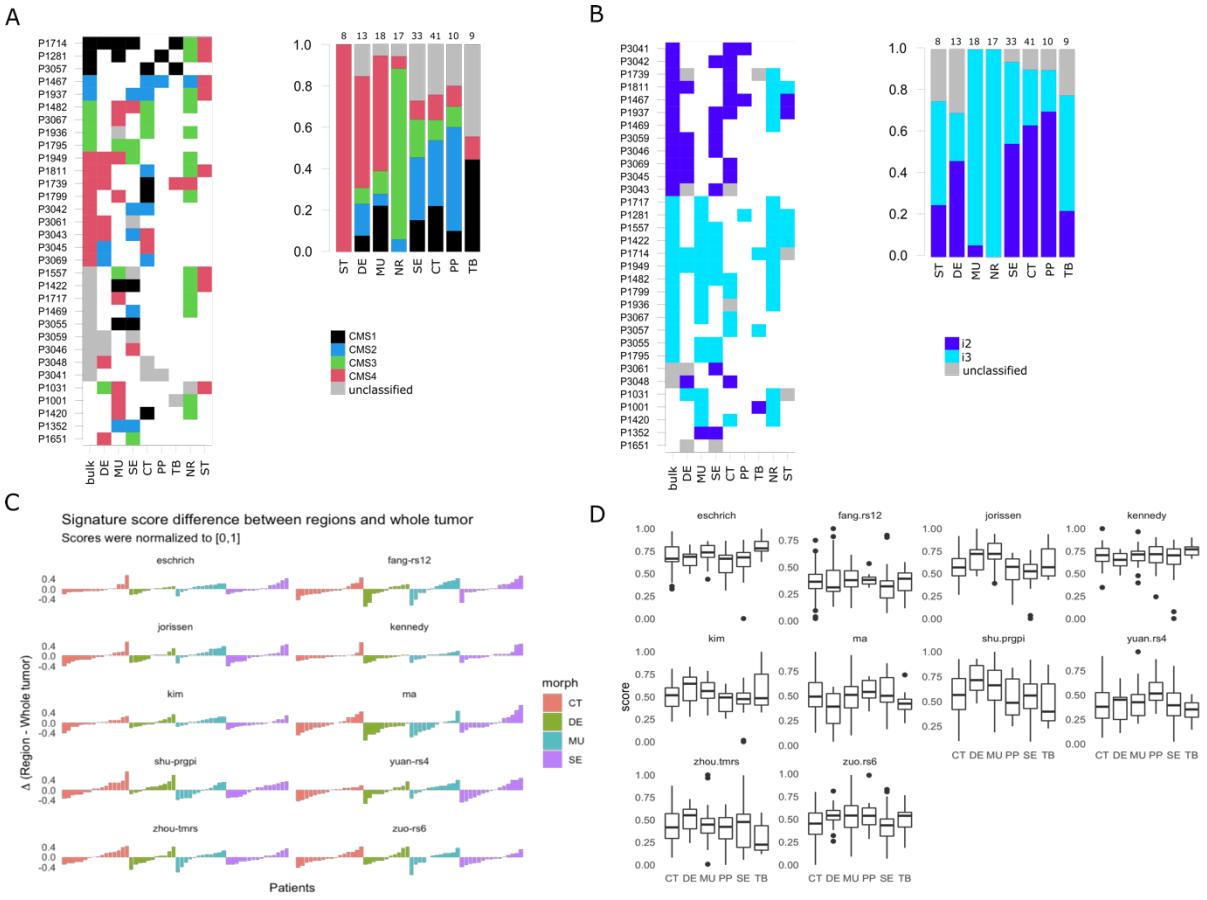
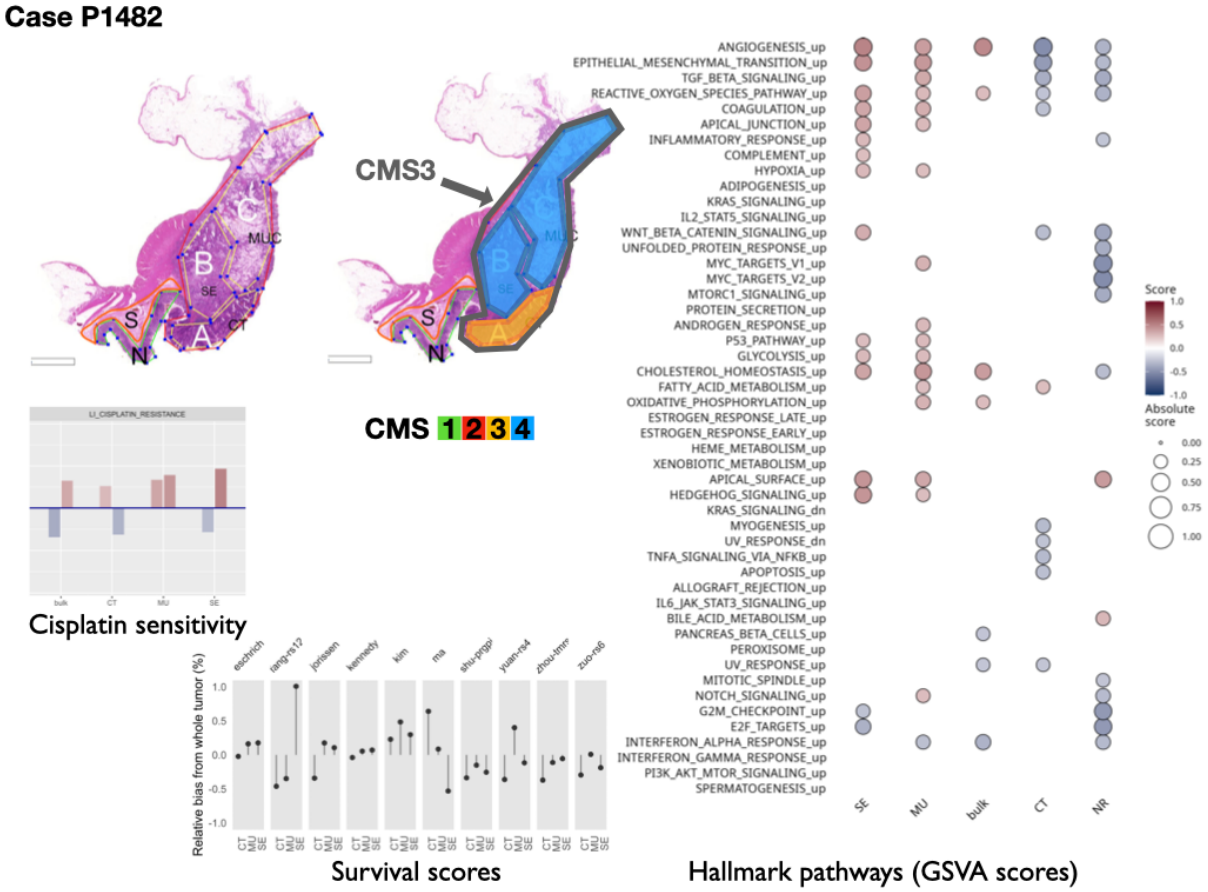


Figure 5

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 Supplemental Figure SF5A-B. 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.