OPEN ACCESS
CellPress

Cell Reports
Resource

Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer

Jesper Bertram Bramsen,[1,11,*] Mads Heilskov Rasmussen,[1] Halit Ongen,[2,3,4] Trine Block Mattesen,[1] Mai-Britt Worm Ørntoft,[1] Sigrid Salling Árnadóttir,[1] Juan Sandoval,[5,10] Teresa Laguna,[5] Søren Vang,[1] Bodil Øster,[1] Philippe Lamy,[1] Mogens Rørbæk Madsen,[8] Søren Laurberg,[9] Manel Esteller,[5,6,7] Emmanouil Theophilos Dermitzakis,[2,3,4] Torben Falck Ørntoft,[1] and Claus Lindbjerg Andersen[1,*]

[1]Department of Molecular Medicine, Aarhus University Hospital, Aarhus 8200, Denmark
[2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva 1211, Switzerland
[3]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva 1211, Switzerland
[4]Swiss Institute of Bioinformatics, Geneva 1211, Switzerland
[5]Cancer Epigenetics and Biology Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet, Barcelona 08908, Catalonia, Spain
[6]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona, Barcelona 08907, Catalonia, Spain
[7]Institucio Catalana de Recerca i Estudis Avancats (ICREA), Barcelona 08010, Catalonia, Spain
[8]Department of Surgery, Hospitalsenheden Vest, Herning 7400, Denmark
[9]Section of Coloproctology, Aarhus University Hospital, Aarhus 8000, Denmark
[10]Present address: Epigenomics Unit, Medical Research Institute La Fe, Valencia 46026, Spain
[11]Lead Contact
*Correspondence: bramsen@clin.au.dk (J.B.B.), cla@clin.au.dk (C.L.A.)
http://dx.doi.org/10.1016/j.celrep.2017.04.045

SUMMARY

Colorectal cancer (CRC) is characterized by major inter-tumor diversity that complicates the prediction of disease and treatment outcomes. Recent efforts help resolve this by sub-classification of CRC into natural molecular subtypes; however, this strategy is not yet able to provide clinicians with improved tools for decision making. We here present an extended framework for CRC stratification that specifically aims to improve patient prognostication. Using transcriptional profiles from 1,100 CRCs, including >300 previously unpublished samples, we identify cancer cell and tumor archetypes and suggest the tumor microenvironment as a major prognostic determinant that can be influenced by the microbiome. Notably, our subtyping strategy allowed identification of archetype-specific prognostic biomarkers that provided information beyond and independent of UICC-TNM staging, MSI status, and consensus molecular subtyping. The results illustrate that our extended subtyping framework, combining subtyping and subtype-specific biomarkers, could contribute to improved patient prognostication and may form a strong basis for future studies.

INTRODUCTION

The prognostication of CRC currently relies on the Tumor Node Metastasis (TNM) staging system. Yet, tumors of the same stage can differ unpredictably in both prognosis and treatment response, which leads to patient under- and overtreatment (Puppa et al., 2010). Several research groups have proposed to resolve this heterogeneity by molecular sub-classification, culminating with the recent proposal of four transcriptional consensus molecular subtypes (CMSs) by the CRC subtyping consortium (CRCSC). The CMSs are associated with distinct histopathological features and it is proposed that molecular subtyping may advance precision diagnostics, treatment, and guide rational drug design (Guinney et al., 2015). However, this remains to be further documented, and consensus molecular subtyping is not yet a tool used to guide clinical decisions. Indeed, further development of molecular stratification approaches may be needed to unveil clinical potentials.

We hypothesized that the major inter-tumor molecular diversity of CRC may have precluded validation of molecular prognostic biomarkers in the past. Relevant biomarkers may well be CRC subtype specific, and the distribution of subtypes in the CRC cohorts used for biomarker identification and validation may have differed.

We therefore pursued that molecular subtyping can establish homogeneous patient subgroups and hereby help to uncover the biological processes associated with aggressiveness within each subtype. Furthermore, molecular subtyping facilitates validation as it enables subtype-specific biomarkers to be validated in tumors of the relevant subtype, rather than in a bulk of molecularly different tumors.

Using this strategy, we here established a framework for CRC prognostication, based on the combination of molecular subtyping and subtype-specific prognostic biomarkers (Figure 1A), which provided prognostic information independently of and beyond CMS subtyping, TNM staging, and microsatellite instability (MSI) status.

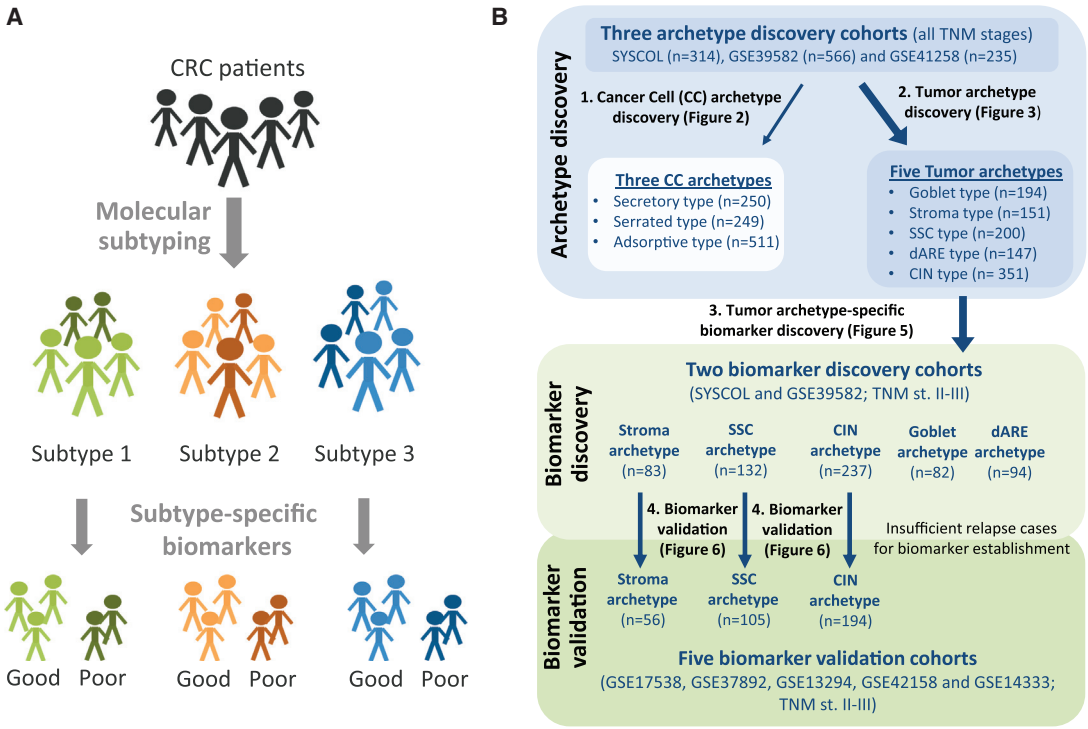


Figure 1. Improved Prognostication of CRC by Molecular-Subtype-Specific Biomarkers

(A) Schematic illustration of the principle of CRC patient prognostication by combining molecular subtyping with subtype-specific prognostic biomarkers. Molecular subtyping is employed to reduce the major inter-tumor molecular diversity of CRC and allows patient prognosis to be more accurately predicted within each subtype by application of subtype-specific prognostic biomarker panels. Patients with good or poor prognosis are indicated.

(B) Overview of the workflow employed for discovery and validation of CRC molecular archetypes (upper panel) and archetype-specific prognostic biomarkers (lower panel). Three cancer cell (CC) and five tumor archetypes were independently identified in three CRC discovery cohorts. Establishment of archetype-specific prognostic biomarkers were based on the stroma, SSC, and CIN tumor archetypes only as the goblet and dARE archetypes contained too few relapse cases for biomarker identification/validation. Prognostic biomarkers were discovered in the SYSCOL and GSE39582 cohorts and validated in samples from the five validation cohorts indicated. Number of samples investigated is given.

RESULTS

Three Cancer Cell Archetypes Exist in CRC

To identify homogeneous molecular archetypes of CRC, we performed RNA sequencing (RNA-seq) and DNA methylation profiling of 33 adenomas and 281 carcinomas (SYSCOL cohort) and unsupervised class discovery by consensus non-negative matrix factorization (NMF)-based clustering using two strategies: one identifying "cancer cell" (CC) archetypes by analyzing only epithelial cell-derived transcripts (as defined from Isella et al., 2015; Figures 1B and 2) and one identifying "tumor" archetypes by analyzing epithelial cell and stroma-derived transcripts together (Figures 1B and 3). The analysis of epithelial cell-derived transcripts and DNA methylation data independently suggested the existence of three distinct CC archetypes (Figures 2A, S1A, and S1B; see Figure S1C for distribution of CC archetypes according to TNM stage, gender, location, and MSI status). Two of the archetypes were named "secretory" and "adsorptive" as gene set enrichment analysis (GSEA) indicated resemblance to secretory and adsorptive enterocyte precursor cell lineages of the normal intestine, respectively (Figure 2B). In agreement, the secretory archetype exhibited key features of secretory cell commitment, such as high *ATOH1* mRNA expression (Figure 2A; Yang et al., 2001), extensive secretory goblet cell differentiation as evaluated by immunohistochemistry (IHC) staining (Figure S1D), and an enrichment for *KRAS* mutations and signaling (Figures 2A and 2D). The adsorptive archetype exhibited classical features of "conventional" adenocarcinomas (Laiho et al., 2007) such as chromosomal instability (CIN; Figures 2A and 2C), microsatellite stability (MSS; Figure 2A), enhanced Wnt/β-catenin-signaling (Figure 2D; Mårtensson et al., 2007), and low DNA methylation (Figures 2A and 2E). The third CC archetype was named "serrated" due to its resemblance to sessile serrated CRC (SSC; Leggett and Whitehall, 2010) including DNA hypermethylation (Figures 2A and 2E), hypermutation, MSI, frequent right-sided location (Figure 2A), and gene expression associated with immune processes, such as interferon (IFN) and inflammatory responses (Figure 2D). For validation of the CC archetypes, we performed independent class discovery in two public CRC datasets GSE39582 (Marisa et al., 2013) and GSE41258 (Sheffer et al., 2009; Figure S1E) and confirmed that the independently predicted archetypes were indeed similar between datasets by principal component analysis (PCA; Figure 2F) and SubMap

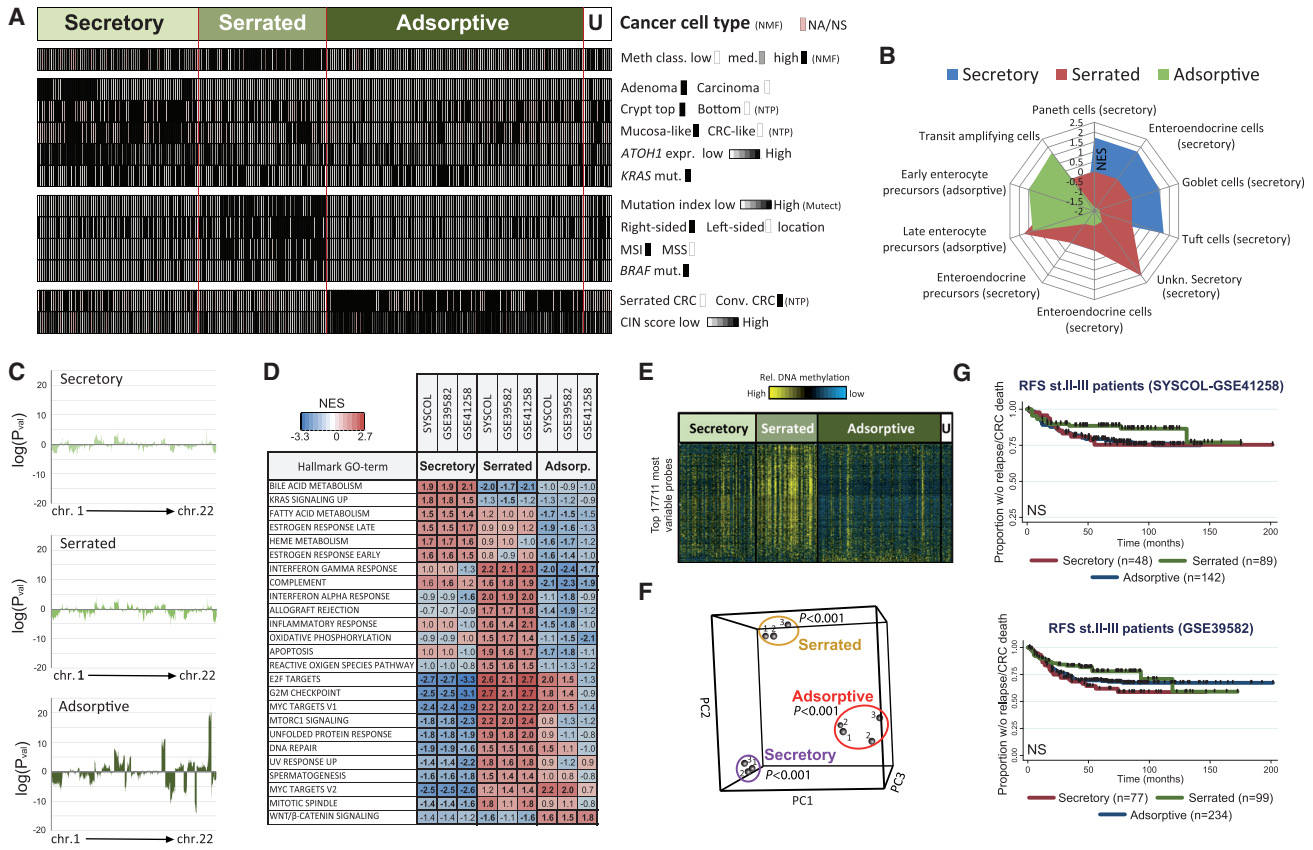


Figure 2. Identification of CC Archetypes

(A) CC archetypes in the SYSCOL cohort identified by consensus NMF-based clustering using epithelial transcripts only. Green indicates CC archetypes, white indicates unclassified samples ("U"), and pink indicates that a value is unavailable (NA) or non-significant (NS). The methylation (Meth) class was determined by independent class discovery using DNA methylation data. NTP analysis was used to compare samples to the published expression signatures as indicated (see Supplemental Experimental Procedures). Mutations and mutation index were extracted from MuTect analysis and CIN scores generated from combined ChAMP/RNA-seq analysis. Scores and expression values are represented as 10 percentile bins.

(B) Comparison of the CC archetypes to normal intestinal epithelial cell types by GSEA. The spider plot shows the archetype-specific normalized enrichment scores (NESs) for ten cell-type-specific gene sets extracted from Grün et al., (2015).

(C) Distribution of copy number alterations (CNAs) along the 22 autosomes stratified according to CC archetypes. Shown is Log_{10} to the p value (two-tailed t test) for comparison of each archetype to normal mucosa. Positive and negative values indicate gains and losses.

(D) Molecular features of the three CC archetypes as evaluated by applying GSEA to the three CRC cohorts SYSCOL, GSE39582, and GSE41258 using "hallmark" gene sets from the Molecular Signatures Database. Gene sets with a positive (i.e., enriched) or negative (i.e., depleted) NES are shown in red and blue and highlighted if all three cohorts were significantly enriched/depleted (FDR <0.1).

(E) Heatmap of DNA methylation levels for the three CC archetypes (median-centered β values for top 17,711 most variable probes of the Infinium HumanMethylation450 BeadChip).

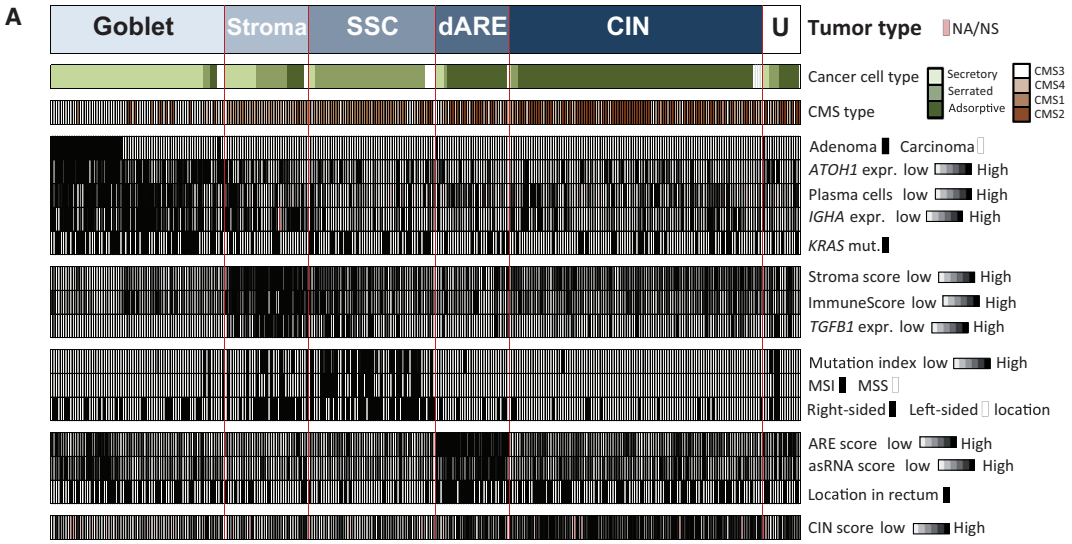
(F) Comparison of independent CC archetype predictions in the SYSCOL, GSE41258, and GSE39582 cohorts by PCA of the CC archetype-specific area under the curve (AUC) scores for genes common to all datasets. Plotted are the first three principal components (PCs), and colored circles represent CC archetypes encompassing a representative from each cohort (black dots; 1 = SYSCOL, 2 = GSE39582, 3 = GSE41258). The adsorptive cluster encompassed two closely related types from the GSE39582 cohort collapsed into one adsorptive archetype during subsequent analysis. Pairwise comparison of the three cohorts by SubMap analysis confirmed that archetypes are similar between cohorts (BH-adjusted p values for each archetype association is given).

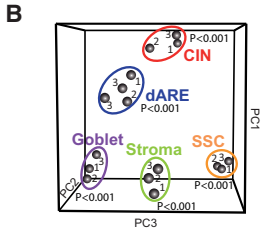
(G) Kaplan-Meier plots showing CC archetype-specific RFS for TNM stage II–III patients in a combined SYSCOL-GSE41258 cohort (top panel) and GSE39582 (lower panel). NS, non-significant (log-rank test).

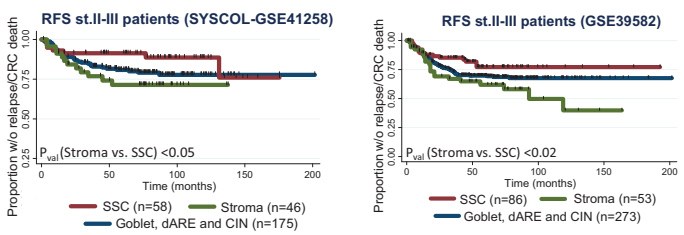
analysis (Figure S1F). Finally, we found no significant difference in relapse-free survival (RFS) between CC archetypes in TNM stage II–III patients in two cohorts, GSE39582 and a combination of SYSCOL-GSE41258 (combined to increase the number of relapse events per archetype; Figure 2G). Hence, the CC archetype appeared not to be a major determinant of patient prognosis per se.

Five Tumor Archetypes Exist in CRC

Although CC archetypes were homogeneous for CC-related traits (Figure 2), we found significant differences in the stroma content of samples within each CC archetype (Figures S1G and S1H), which may relate to patient prognosis. Therefore, to integrate the contribution of the tumor stroma into CRC archetypes, we performed class discovery in the SYSCOL cohort



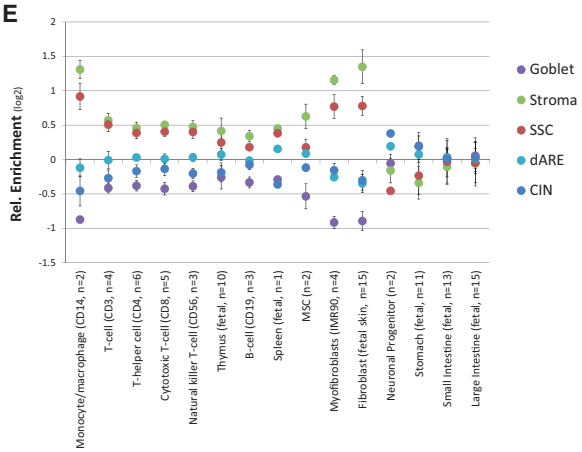


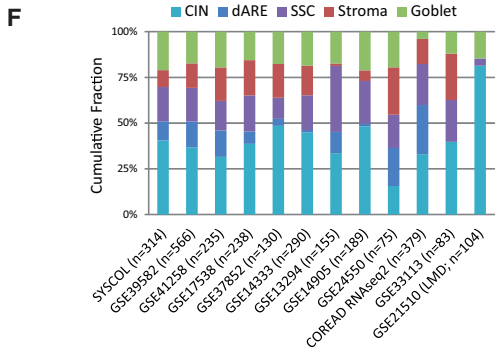


D

NES
-3.5 0 4.2

	Gene-set	Goblet SYSCOL	Goblet GSE39582	Goblet GSE41258	Stroma SYSCOL	Stroma GSE39582	Stroma GSE41258	SSC SYSCOL	SSC GSE39582	SSC GSE41258	dARE SYSCOL	dARE GSE39582	dARE GSE41258	CIN SYSCOL	CIN GSE39582	CIN GSE41258
CRC	CRC vs. Mucosa[1]	-2.3	-2.5	-2.2	-2.5	-2.4	-2.0	1.7	1.3	1.7	1.0	-1.1	-0.6	2.4	2.9	3.1
	E2F targets[2]	-2.6	-2.8	-2.6	-2.8	-3.5	-3.1	2.6	2.4	3.3	-1.4	-1.6	-2.1	2.5	2.9	3.0
	G2M checkpoint[2]	-2.3	-2.6	-2.5	-2.2	-2.8	-2.6	2.7	2.2	3.0	-1.6	-1.7	-1.8	2.1	2.6	2.8
	MYC sign v2[2]	-2.2	-2.4	-2.0	-2.9	-3.0	-2.7	1.6	1.7	2.5	-0.5	1.1	-0.5	2.5	2.2	2.8
	MTORC1 sign.[2]	-2.0	-1.7	-2.2	0.9	-1.5	-1.1	2.5	2.0	2.5	-1.6	-1.2	-1.6	1.4	-1.0	1.9
Stroma	Normal stroma[7]	-1.1	-0.4	-2.1	3.0	3.4	3.5	-1.9	-2.4	-2.4	-1.6	-1.6	-1.1	0.8	-2.0	-2.1
	Activated stroma[7]	-2.3	-1.8	-3.5	4.2	4.0	4.2	0.6	-1.6	-1.0	-1.6	-0.7	0.8	0.6	-3.4	-2.9
	EMT[8]	-2.8	-2.3	-3.4	3.5	3.3	3.6	0.9	-3.0	-1.8	-0.9	1.4	1.7	1.4	-2.8	-2.5
	CAF[3]	-2.4	-1.6	-2.9	3.3	3.2	3.3	-1.7	-3.0	-2.2	-1.2	1.2	1.3	1.4	-2.5	-2.0
	MSC[5]	-1.3	1.1	-2.1	3.5	3.1	3.1	-1.0	-1.4	-1.4	-1.3	-0.8	1.2	-1.7	-2.9	-2.6
	Myofibroblast[5]	-1.1	1.2	-1.8	3.3	2.9	2.8	-1.4	-1.7	-1.7	-1.2	-0.6	1.7	-1.6	-2.8	-2.6
	Myogenesis[2]	-1.3	-0.8	-1.9	2.5	2.6	2.6	-1.7	-2.4	-2.1	-0.6	1.8	1.8	-1.0	-2.2	-2.6
	Endothelial[3]	-1.8	-0.9	-2.1	2.6	2.3	2.4	-1.7	-1.0	-1.1	-1.4	-0.8	0.7	1.4	-2.1	-1.6
Immune cells	Leucocytes[3]	-1.1	1.3	-1.3	2.7	2.1	2.0	-0.6	0.8	0.9	-1.5	0.5	0.8	-1.8	-2.7	-2.4
	Macrophages[5]	-1.2	-0.9	-1.6	1.9	1.9	2.0	-1.5	-1.6	-1.1	0.7	1.1	-0.7	-0.8	-1.9	-1.5
	B-cells, activated[5]	-1.1	1.3	-0.8	2.3	1.9	1.6	-1.1	0.9	0.6	-0.9	0.7	0.8	-1.6	-2.2	-1.8
	B-cells, immature[5]	1.0	1.3	-0.4	1.9	1.8	1.7	0.9	-0.6	0.9	-1.4	-0.8	-1.0	-1.6	-1.7	-1.9
	DC[5]	-1.6	-1.2	-1.9	2.4	2.3	2.3	-0.6	-0.6	0.6	-1.3	-1.0	1.0	-0.5	-2.0	-1.9
	NK cells[5]	-1.2	-1.3	-1.6	2.1	2.0	1.9	-0.7	-0.8	-0.6	-1.2	0.6	0.9	0.7	-1.9	-1.8
	Neutrophil cells[5]	-1.7	-1.2	-1.5	1.7	1.1	1.2	1.2	1.1	1.2	0.7	1.1	0.9	0.7	-1.3	-0.9
	T-cells[5]	-1.5	1.2	-1.4	3.0	2.4	2.2	0.9	1.5	1.6	-1.5	-0.5	1.4	-1.8	-2.7	-2.6
	CD4 T-cells act.[5]	-2.1	-2.1	-2.0	-1.8	-2.6	-2.0	2.5	2.4	2.8	-1.4	-1.7	-1.8	1.6	2.1	2.0
	CD8 T-cells act.[5]	-1.7	-2.0	-1.6	-1.8	-2.4	-2.2	1.7	1.6	2.2	-0.8	-1.0	-1.7	1.8	1.9	2.2
	T-Effector memory CD8[5]	-0.9	1.1	-1.6	2.5	2.3	2.5	1.6	1.5	1.9	-1.7	-1.3	-1.4	-2.3	-2.8	-2.5
Immune process	IFN alpha sign.[2]	-1.4	-0.6	-1.8	2.3	1.8	2.1	1.9	1.9	2.3	-0.5	-1.1	-1.2	-2.0	-2.3	-2.0
	IFN gamma sign.[2]	-1.5	0.7	-1.9	3.1	2.3	2.4	2.1	1.8	2.5	-1.1	-1.2	-1.4	-2.3	-3.0	-2.5
	Allograft rejection[2]	-1.4	-0.5	-1.5	2.7	2.3	2.2	1.4	1.3	1.5	-1.4	-1.2	-0.9	-1.5	-2.7	-2.5
	Complement[2]	-1.4	0.9	-1.7	2.8	2.3	2.5	1.6	1.3	1.5	-1.4	-1.0	-1.5	-1.7	-2.6	-2.1
	Inflammatory reponse[2]	-1.3	0.9	-1.7	3.1	2.3	2.5	1.4	1.2	1.3	-1.7	-1.1	-0.7	-2.0	-2.8	-2.5
Immune-suppr.	Central memory CD8[5]	-1.2	-0.8	-1.6	2.1	2.2	2.1	-0.8	-1.1	-0.8	-1.5	-1.1	-0.9	-0.6	-1.6	-1.9
	Treg[5]	-1.9	-1.0	-2.3	2.3	2.2	2.1	-0.6	-1.2	-0.7	-1.3	0.8	1.1	0.9	-1.9	-2.2
	MDSC[5]	-1.0	0.7	-1.4	2.2	1.9	2.0	-0.8	-0.7	-1.0	-1.2	0.9	1.1	0.7	-1.8	-1.9
	TGF beta sign.[2]	-0.8	-0.8	-1.1	1.9	2.0	1.9	-1.0	-0.9	-1.1	-1.5	-1.2	-1.2	0.5	-0.9	-0.7





(legend on next page)

using transcripts of both epithelial and stromal origin. This identified five tumor archetypes, denoted "goblet," "stroma," "SSC," "dARE," and "CIN" (Figures 3A and S2A; see Figure S2B for distribution of tumor archetypes according to TNM stage, gender, location, and MSI status), which were validated in the two independent CRC cohorts as described above for the CC archetypes (Figures 3B, S2C, S2D, S2E, and S2F). The goblet and SSC archetypes were dominated by the secretory and serrated CC archetypes, respectively (Figure 3A) and characterized by the key features presented for those above (Figure 2). More notably, the stroma tumor archetype encompassed a mix of all three CC archetypes (Figure 3A) and was best characterized by properties of the tumor microenvironment (TME). Stroma tumors showed high expression of transcripts derived from the tumor stroma (Figure S2G) that were predicted to have both immune cell and non-immune cell origin by the ESTIMATE tool (Figure 3A; "ImmuneScore" and "Stroma score"). Finally, the inclusion of stromal transcripts identified two tumor archetypes, "dARE" and "CIN" that both belonged to the adsorptive CC archetype and shared features characteristic of conventional CRC such as CIN (Figure 3A). Notably, we found that patients with stroma archetype tumors had shorter RFS than other patients, particularly the SSC tumor patients, in the GSE39582 and SYSCOL-GSE41258 cohorts (Figures 3C and S2H). This indicated that the TME had a stronger impact on patient prognosis than the CC archetypes (Figure 2G). To identify the biological processes that distinguished tumor archetypes, and possibly affected patient prognosis, we therefore performed GSEA focusing on TME-related traits. Foremost, both the poor-prognosis stroma tumors and good-prognosis SSC tumors were enriched in gene sets associated with "immune processes" and exhibited high T cell infiltration as evaluated by both RNA-seq (Figure 3D) and DNA methylation profiling (Figure 3E). However, SSC tumors were enriched in transcripts defining active cytotoxic CD4 and CD8 T cells and depleted in transcripts associated with an activated "stroma" and "immune suppression," whereas the reverse pattern was observed in poor-prognosis stroma tumors (Figure 3D). A relative enrichment of stromal cells, such as myofibroblasts (IMR90) and mesenchymal stem cells (MSCs) in stroma tumors was observed by RNA-seq (Figure 3D) and confirmed by DNA methylation analysis (Figure 3E). The CIN and dARE tumors were instead relatively depleted in gene sets associated with both stromal and immune activity, in particular, IFN-α/γ signaling and T cells, again seen in both RNA-seq (Figure 3D) and DNA methylation data (Figure 3E). Finally, the goblet tumor TME was most similar to normal mucosa as evaluated by PCA of stromal transcripts (Figure S3A) and characterized by high expression of immunoglobulin A (*IGHA1* and *IGHA2*), the principal antibody of normal intestine (Figure 3A).

Interestingly, we found that the tumor archetype distribution differed between public CRC cohorts (Figure 3F), which had been subtyped into tumor archetypes using our "CRCclassifier" (Figure S3B). For example, the GSE13294 dataset contained relatively many SSC tumors, whereas goblet tumors were more infrequent in the TCGA colorectal samples (COREAD). Notably, the stroma, dARE and SSC tumor archetypes, which are in part characterized by stromal transcripts, were diminished or completely absent in the laser capture microdissected (LMD) cohort (GSE21510; Figure 3F). We also evaluated how the sampling of spatially distinct regions of the tumor impacted archetype assignment (Figure S3C). Archetype assignments were overall very robust and only the stroma archetype assignment varied in one biopsy from one patient, which may be expected given some variance in the stromal content between tumor biopsies. Collectively, our results suggest that properties of the TME, primarily anti-tumor immune cell and fibroblast activity, are associated with the observed differences in patient prognosis. However, our results also suggest that the evaluation of such TME-related traits may, at least in some cases, be influenced by the biopsy site and the chosen criteria for cancer cell content.

Figure 3. Identification of Tumor Archetypes

(A) Tumor archetypes in the SYSCOL cohort identified by consensus NMF-based clustering using all transcripts. Blue, green, and brown indicate the tumor archetype, CC archetype, and the predicted CMS consensus type (CMS1–4). White indicates unclassified samples ("U"), and pink indicates that a value is unavailable (NA) or non-significant (NS). The plasma cell content was evaluated by CIBERSORT. *IGHA* expression is the summed expression of *IGHA1* and *IGHA2*. Stroma score and ImmuneScore were evaluated using the ESTIMATE package. The ARE score represents the mean expression of ARE-negative divided by the mean expression of ARE-positive transcripts (max-min-normalized), whereas the asRNA score represents the average max-min-normalized expression of asRNA transcripts. Scores and expression values are represented as 10 percentile bins.

(B) Comparison of independent NMF-based tumor archetype predictions in the SYSCOL, GSE41258, and GSE39582 cohorts. Plotted are the first three PCs from a PCA of the archetype-specific AUC values for each gene. Colored circles represent tumor archetypes each containing a representative archetype from each cohort (black dots; 1 = SYSCOL, 2 = GSE39582, 3 = GSE41258). The dARE cluster encompassed two closely related types from the GSE41258 cohort collapsed into one dARE archetype during subsequent analysis. Pairwise comparison of the three cohorts by SubMap analysis confirmed that archetypes are similar between cohorts (BH-adjusted p values for each archetype association are given).

(C) Kaplan-Meier plots showing tumor-archetype-specific RFS for TNM stage II–III patients in a combined SYSCOL-GSE41258 cohort (left panel) and GSE39582 (right panel). Goblet, dARE, and CIN samples have been collapsed into one group for illustration purposes (see Figure S2H for non-collapsed data). p values (log-rank test) are given for SSC versus stroma tumor groups.

(D) Molecular features of the tumor archetypes as evaluated by GSEA in the SYSCOL, GSE41258, and GSE39582 cohorts. Features with a positive (i.e., enriched) or negative (i.e., depleted) NES are shown in red and blue, respectively, and in bold if FDR <0.1. The gene sets underlying each feature were extracted from the sources: [1]Grade et al. (2007) (MSigDB:M16740); [2]MSigDB H: hallmark gene sets; [3]Isella et al. (2015); [5]Angelova et al. (2015); [7]Moffitt et al. (2015); [8]Anastassiou et al. (2011) (MSigDB: M2572); and [9]Harris (2002) (MSigDB:M10508).

(E) DNA methylation profiles of gene promoter regions were used to estimate the relative enrichment/depletion of a range of cell types in the five tumor archetypes using the eFORGE tool. Error bars indicate the SD.

(F) Tumor archetype distribution of publicly available CRC transcriptome datasets as determined by our CRCclassifier. GEO accession numbers and sample size for the datasets are given. LMD, laser capture microdissected.

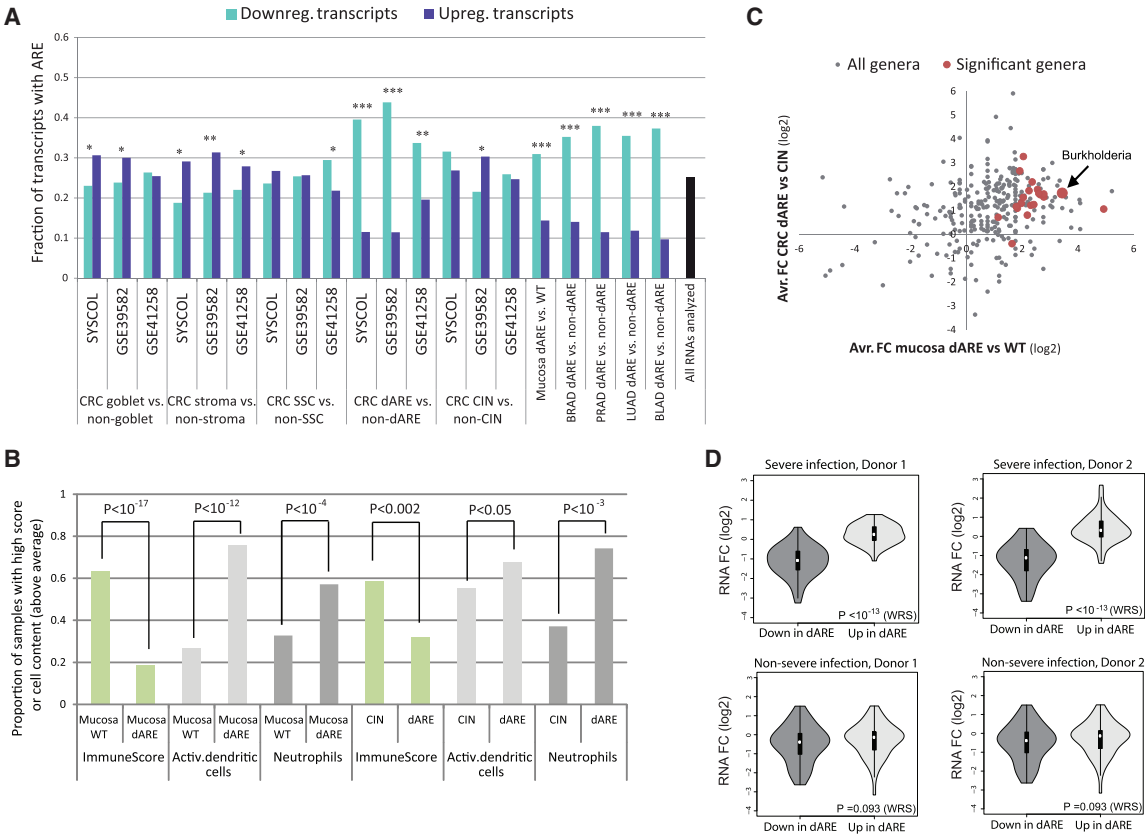


Figure 4. The dARE Archetype Is Influenced by the Microbiome

(A) The fraction of ARE-positive transcripts among all up- and downregulated transcripts (median FC >1 and FC <1, respectively; both $p < 0.05$; WRS) in the dARE archetype of CRC cohorts (SYSCOL, GSE39582, GSE41258), normal mucosa samples (SYSCOL), breast cancer (BRAD; TCGA), prostate cancer (PRAD; TCGA), lung cancer (LUAD; TCGA), and bladder cancer (BLAD; TCGA). "All RNAs analyzed" indicate the fraction of ARE-positive transcripts among all transcripts included in the analysis (i.e., the "background" proportion of ARE-positive transcripts). An asterisk (*) indicates that the distribution of ARE-positive transcripts within a dataset is significantly different between upregulated and downregulated transcripts at the following significance levels: *$p < 0.05$; **$p < 10^{-7}$; ***$p < 10^{-20}$ (WRS).

(B) Proportion of samples within the mucosa WT, mucosa dARE, CIN, and dARE tumor archetypes with a high (i.e., above average) ImmuneScore (evaluated by the ESTIMATE package) or content of activated DCs and neutrophils (as evaluated by the CellMix tool). Significance was assessed using WRS.

(C) The average fold change (FC) in bacteria genera read numbers comparing mucosa dARE to mucosa WT (x axis) and CRC dARE to CIN archetypes (y axis). Bacteria genera with significantly higher read counts in both CRC and mucosa dARE are shown in red ($p < 0.01$; WRS).

(D) Violin plot showing the RNA expression FC of the top 50 RNA transcripts upregulated or downregulated in the dARE versus CIN tumor archetypes of the three discovery cohorts SYSCOL, GSE39582, and GSE41258 in DCs from two healthy donors after stimulation by plasma severely or non-severely infected with the bacteria Burkholderia (data from GSE49753). Median FC (white dot), interquartile range (black boxes), 95% confidence interval (black line), and statistical significance (WRS) are indicated.

The dARE Archetype Is Microbiome Dependent

We found an overall good resemblance between our tumor archetypes and the four proposed CMS subtypes upon application of the CMS classifier provided by the CRCSC (Guinney et al., 2015). A very notable exception, however, was our further stratification of CMS2 into the dARE and CIN tumor archetypes (Figures 3A and S4A). Interestingly, dARE tumors were depleted in transcripts containing 3′ UTR AU-rich elements (AREs), which motivated the name depleted in AU-rich elements (dARE) and enriched in antisense RNA (asRNA)/long non-coding RNA (lncRNA; Figures 3A and 4A and S2G). We observed a similar dARE-like phenotype characterized by depletion of ARE-positive (ARE^+) transcripts in a group of 70 (of 301) normal mucosa samples identified by NMF-based clustering (referred to as "mucosa dARE" opposed to "mucosa wild-type (WT)"; Figures 4A, S4B, and S4C). AREs are frequently found in immune related transcripts, where they facilitate the post-transcriptional degradation upon stimulation, e.g., by anti-inflammatory interleukin-10 (IL-10; Kishore et al., 1999). In accordance, we found reduced expression of immune transcripts in both tumor and mucosa dARE samples as compared to CIN tumors and mucosa WT samples, respectively (ImmuneScore; Figure 4B). Colonic dendritic cells (DCs) can be activated to produce IL-10 upon infection by certain bacteria and attract neutrophilic cells (Rigby et al., 2005; Doz et al., 2013). Indeed, the mucosa and tumor dARE groups encompassed significantly more samples with a high content of activated DCs ($p < 10^{-12}$ and $p < 0.05$; Wilcoxon rank-sum [WRS] test) and neutrophils ($p < 10^{-4}$ and $p < 10^{-3}$; WRS) than CIN samples as evaluated by the CellMix tool (Figure 4B). To investigate whether bacterial infection may have induced DC activation in

dARE samples, we mapped SYSCOL RNA-seq reads to bacterial genomes. We found that the proportion of samples with a high bacterial read count was significantly higher for dARE than CIN tumor samples ($p < 2 \times 10^{-7}$; WRS), for mucosa dARE than mucosa WT samples ($p < 3 \times 10^{-5}$; WRS; Figure S4D) and that the bacteria genus Burkholderia was enriched in both mucosa ($p < 3.3 \times 10^{-3}$; WRS) and tumor ($p < 7.3 \times 10^{-4}$; WRS) dARE samples (Figure 4C). To investigate whether Burkholderia infection could induce a dARE-like transcriptional phenotype in DCs, we analyzed data from a study where DCs from two healthy donors were exposed to Burkholderia-infected plasma (Khaenam et al., 2014; GSE49753). Indeed, the exposure to severely infected plasma induced a dARE-like phenotype with overrepresentation of ARE^{+} transcripts among downregulated genes (Figure S4E) similar to our clinical samples (Figure 4A). In agreement, the majority of top 50 up- or downregulated transcripts in dARE tumors were similarly up- and downregulated in DCs from two healthy donors upon exposure to severely infected Burkholderia plasma ($p < 10^{-13}$; WRS; Figure 4D). Furthermore, we found dARE-like transcriptional profiles upon class discovery in other cancers including breast, prostate, bladder, and lung cancer (Figure 4A), suggesting a possible microbial influence on a subfraction of these cancers, which require further investigation.

Identification of Archetype-Specific Prognostic Biomarkers

The observation that the TME was strikingly different between tumor archetypes and related to patient prognosis suggested that prognostic biomarkers may well be archetype specific rather than universal. Therefore, we investigated whether subtyping of CRC into homogeneous tumor archetypes would allow identification and validation of archetype-specific prognostic biomarkers (Figure 1). We divided the TNM stage II–III tumors from the SYSCOL and GSE39582 cohorts into aggressive (CRC relapse/CRC-related death) and non-aggressive groups (no relapse or CRC-related death) within the stroma, SSC, and CIN archetypes and compared aggressive and non-aggressive samples by GSEA (the goblet and dARE archetype tumors were excluded from this analysis due to low numbers of relapsing cases). Overall, we found that particularly gene sets related to immune processes (e.g., lymphocyte activation, immune processes and IFN responses) were depleted in aggressive tumors for all three tumor archetypes (Figure 5A; see Figure S5A for top enriched/depleted gene sets distinguishing aggressive and non-aggressive tumors in SYSCOL/GSE39582 archetypes). Conversely, gene sets related to an activated stroma/EMT (epithelial-mesenchymal transition) (i.e., ECM/EMT) were enriched in aggressive stroma and SSC tumors, whereas gene sets associated with respiratory electron transport (RET)/oxidative phosphorylation (Oxphos) were depleted in aggressive CIN tumors for both discovery cohorts (Figure 5A). This suggested that the same biological processes (e.g., immune cell activity and stromal EMT processes) distinguished aggressive and non-aggressive tumors of the SSC and stroma archetypes. However, inspection of the top-enriched gene set "mesenchymal transition signature" (Anastassiou et al., 2011) in the SSC and stroma archetypes revealed that the particular transcripts driving the EMT gene-set enrichment were archetype specific. Most EMT-associated transcripts with a high ranked metric score (RMS; i.e., enrichment) in the stroma archetype had a low RMS in the SSC archetype and vice versa (Figure 5B). In agreement, a biomarker panel based on the top ten enriched/depleted EMT-related transcripts in aggressive stroma tumors were prognostic only in stroma tumors and not SSC tumors, and vice versa (Figures 5C and S5B). This was validated in the independent cohort GSE17538 (Figure 5D). We next focused on aggressive CIN tumors where top depleted gene sets were associated with RET (Figures 5A, 5E, and S5A). Similarly, a panel of the top ten enriched/depleted RET-related transcripts was only prognostic in CIN tumors and not in SSC/stroma tumors (shown for the validation cohort GSE17538 in Figure 5F and for the discovery cohorts SYSCOL and GSE39582 in Figure S5C). Collectively, these observations show that archetype-specific prognostic biomarkers exist and underscore the strength of our strategy to molecularly subtype CRC cohorts prior to biomarker validation attempts.

Establishment of Composite Prognostic Biomarker Panels for CRC

Our above analysis indicated that several biological traits contributed to aggressiveness within each tumor archetype (such as immune signaling, EMT processes, and RET/oxphos; Figure 5A) and that archetype-specific biomarker panels for each of these traits could be established. To further improve the prognostication of CRC patients, we therefore next generated "composite" biomarker panels that interrogate several aggressive traits per archetype by summing the biomarker scores for the individual traits. For stroma tumors, we combined three biomarker panels for enhanced EMT, DNA methylation loss, and reduced IFN-γ signaling, for SSC tumors, we combined three biomarker panels for enhanced EMT, stromal stem cell (STC) activity, and reduced IFN-γ signaling and for CIN tumors four biomarker panels for reduced RET/Oxphos/peripheral blood mononuclear cell (PBMC)/IFN-γ signaling were combined (for details on how the composite panels were derived see Supplemental Experimental Procedures section "Establishment of archetype-specific prognostic biomarker panels and calculation of panel scores"; the relevant gene sets are marked with asterisk in Figure 5A). Each of these three composite biomarker panels generated a panel score (P-score), which were denoted P-stroma, P-SSC, and P-CIN for the stroma, SSC, and CIN archetypes, respectively. We confirmed that our three archetype-specific composite panels stratified patients from the stroma, SSC, and CIN archetypes into groups with significantly different RFS in the discovery cohorts SYSCOL and GSE39582 (Figure 6A; left panels). We next validated the P-scores in samples from independent public cohorts sufficiently sized for archetype-specific analysis of RFS (Figure 6A, right panels; cohorts were combined to increase number of relapse events in each cohort; see Supplemental Experimental Procedures section "Relapse-free survival analysis and samples included"): here, the CIN panel could be evaluated in more validation cohorts than the SSC and stroma panels due to the higher frequency of the CIN archetype. Also, we confirmed that the archetype-specific panels were prognostic only in the intended tumor archetype in both discovery and validation cohorts (Figure S6A). We next compared the hazard ratios (HRs) and HR 95% confidence interval (HR95%CI) for P-scores and TNM stage

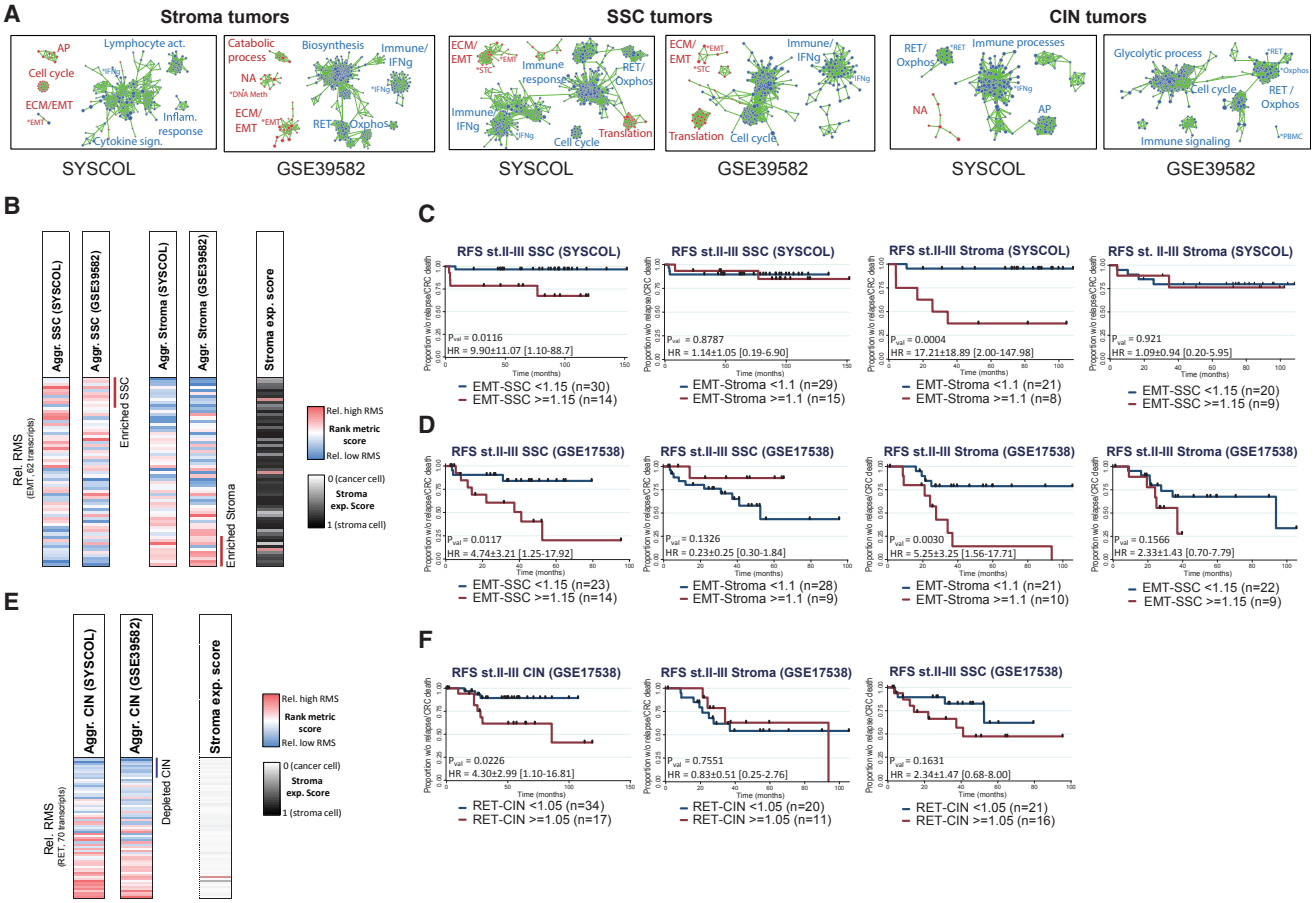


Figure 5. Identification of Subtype-Specific Prognostic Biomarker Panels

(A) GSEA enrichment maps showing gene sets enriched (red circles) or depleted (blue circles; connected by green lines) in aggressive stroma, SSC, and CIN archetype tumors from the SYSCOL and GSE39582 cohorts (FDR <0.1). Text labels indicate the most prominent gene ontology features associated with the particular gene-set cluster. ECM, extracellular matrix; AP, antigen presentation; EMT, epithelial-to-mesenchymal transition; STC, stroma stem cells; RET, respiratory electron transport; Oxphos, oxidative phospohorylation; INFg, IFN-γ signaling; PBMC, peripheral blood mononuclear cell; DNA meth., DNA methylation loss; UN, unknown function. Gene sets utilized for establishment of archetype-specific prognostic biomarker panels are indicated with asterisks (*).

(B) Distribution of the GSEA rank metric scores (RMS) for transcripts of the gene set mesenchymal transition signature (EMT; Anastassiou et al., 2011; MSigDB: M2572) comparing aggressive and non-aggressive SSC and stroma tumors in the SYSCOL and GSE39582 datasets. Red indicates a high relative RMS and that the transcript is enriched in aggressive tumors. Top transcripts enriched (i.e., high RMS) in aggressive tumors of the stroma and SSC archetypes are indicated with red bars. The stroma expression (expr.) score (fraction of transcripts of stromal origin; Isella et al., 2015) is given to the right.

(C and D) Kaplan-Meier survival plots showing the RFS of TNM stage II–III SSC (left two panels) and stroma (right two panels) patients of the of SYSCOL (C) and GSE17538 (D) cohorts stratified by the EMT-SSC and EMT-stroma prognostic biomarker panels as indicated. p values (log-rank test) and HR95%CI are indicated.

(E) Distribution of the GSEA rank metric scores (RMS) for transcripts of the gene set respiratory electron transport, ATP synthesis by chemiosmosis coupling, and heat production by uncoupling proteins (here denoted RET; MSigDB: M1025) comparing aggressive and non-aggressive CIN tumors in the SYSCOL and GSE39582 cohorts. Top transcripts depleted (i.e., low RMS) in aggressive CIN tumors are indicated with a blue bar. The stroma expr. score is given to the right.

(F) Kaplan-Meier survival plots showing the RFS of CIN, stroma, and SSC TNM stage II–III patients of the GSE17538 cohort stratified by the RET-CIN prognostic biomarker panel. p values (log-rank test) and HR95%CI are indicated.

in a univariate (Figure 6B) and multivariate Cox regression analysis (Figure S6B). Multivariate Cox regression analysis, also including the molecular subtypes (CMS or tumor archetype), showed that the P-scores were independent predictors of RFS and overall stronger than both TNM stage and molecular subtypes (Figures 6C and S6C). The P-score added significant prognostic value to both TNM stage II and III patients (stage II: HR, 5.06; 95% CI, 3.24–7.92; p = 0; Figure 6D, center panel; stage III: HR, 4.05; 95% CI, 2.67–6.13; p = 0; Figure 6D, right panel). We applied Harrell's C-index to estimate how much the predictive accuracy of the regression model was increased by adding the P-score to the TNM stage. A C-index increase of 0.14 and 0.06 was observed in the discovery and validation cohorts, respectively, confirming that the P-score improved the predictive accuracy. Finally, the P-score also provided prognostic information additional to MSI status, a well-established marker of favorable prognosis (Saridaki et al., 2014), in the patient samples for which tumor MSI/MSS status was available (Figure S6D).

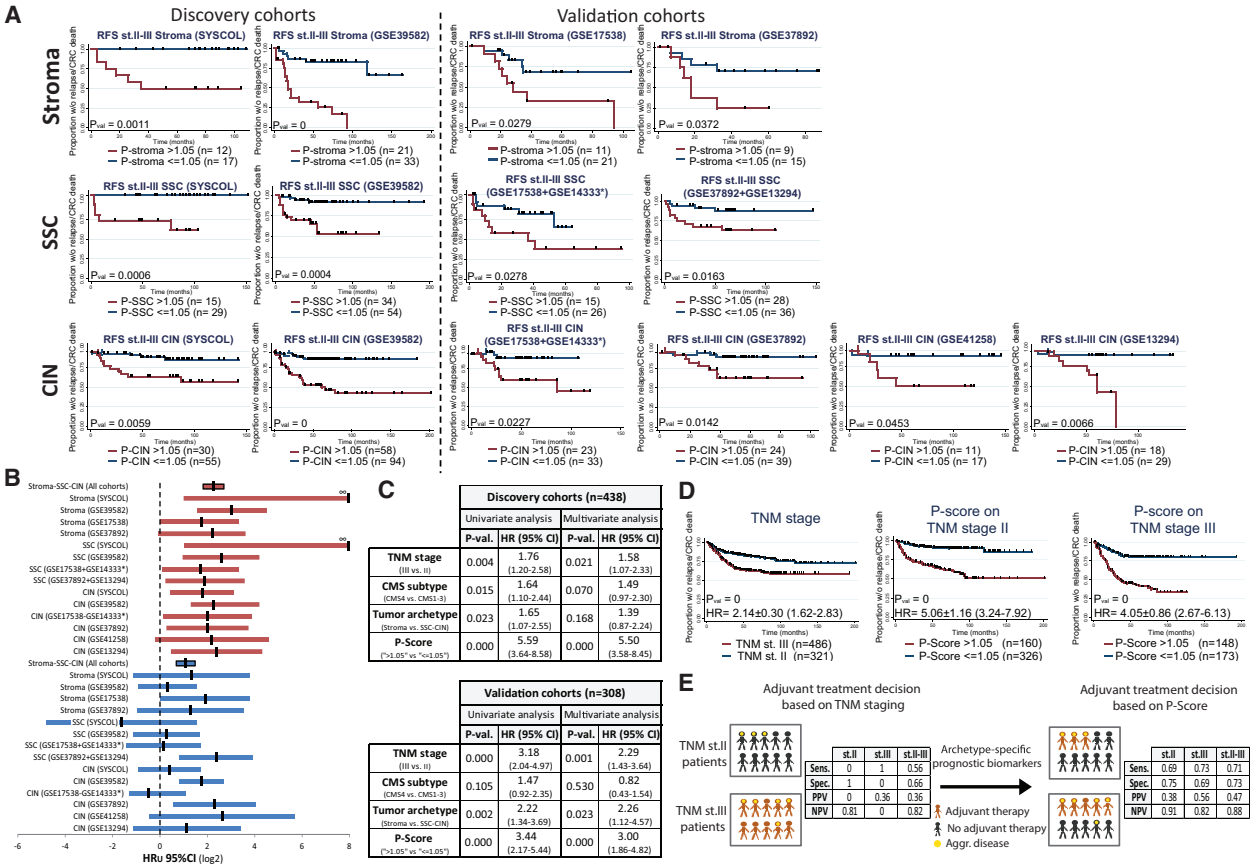


Figure 6. Validation of Composite Prognostic Biomarker Panels

(A) Kaplan-Meier survival plots showing the RFS of stroma (top panel), SSC (middle panel), and CIN (lower panel) tumor patients stratified by the archetype-specific prognostic biomarker panels P-stroma, P-SSC, and P-CIN. Plots for the two prognosis discovery cohorts, SYSCOL and GSE39582, are shown in the left panel, whereas independent validation cohorts are shown in the right panel. See Supplemental Experimental Procedures for a description of the cohorts used. p values are indicated (log-rank test).

(B) Forest plot of the HRs (black lines) and HR95%CIs estimated for the prognostic biomarker panels (red) and tumor TNM stage (blue) in CRC cohorts evaluated by a univariate Cox regression analysis. ∞ indicates that the HR95%CI has an infinite value and was here set to 1–8 for illustration purposes.

(C) Tables showing the significance and HR for TNM stage, CMS subtype, tumor archetype, and P-score in CRC samples of the discovery cohorts (upper table) and validation cohorts (lower table) included in (A) using a univariate and multivariate cox regression analysis (TNM stage, CMS subtype, tumor archetype, and P-scores were co-variables). The analysis was restricted to CRC samples for which a CMS annotation was provided by the CRCSC (Guinney et al., 2015) or calculated using the CMSclassifier for the SYSCOL/GSE41258 cohorts, which were not analyzed by the CRCSC.

(D) Kaplan-Meier survival plot showing the RFS of stroma, CIN, and SSC TNM stage II–III tumor patients from all cohorts analyzed in (A) stratified by TNM stage only (left panel) or by P-scores in TNM stage II (middle panel) and III tumors (right panel). p values (log-rank test), HR (±SD), and HR95%CI are given for a univariate cox regression analysis (HR).

(E) Schematic illustration of the potential application of P-scores to stratify TNM stage II–III patients with aggressive disease for adjuvant chemotherapeutic treatment. Typically, adjuvant chemotherapy is offered to stage III patients only leading to systematic under- and overtreatment of stage II and stage III patients, respectively (left panel). The P-score helps reduce this problem by identifying patients with aggressive disease within the stage II and stage III patient groups, which allows targeting of chemotherapeutic treatment to these patients only. The sensitivity, specificity, NPV, and positive predictive value (PPV) for the TNM stage (left) and P-score (right) in TNM stage II and III patients included in (A) are given.

Collectively, this illustrated that our archetype-specific biomarker panels enable CRC prognostication to be enhanced beyond CMS subtyping, MSI status, and the routinely used TNM staging.

DISCUSSION

In this work, we hypothesized that molecular stratification of CRC into homogeneous subtypes is required for validation of robust prognostic biomarkers across CRC cohorts. We therefore initially identified molecular archetypes of CRC by performing class discovery in independent CRC cohorts, both upon exclusion and inclusion of stromal transcripts. This approach adds to previous subtyping strategies that have not weighed the cellular origin of the analyzed transcripts (Guinney et al., 2015) and allowed us to propose a model of CRC in which three major CC archetypes are encompassed within a total of five major tumor archetypes.

The three CC archetypes agree with the proposal of three molecular origins of CRC (Leggett and Whitehall, 2010). Here, the adsorptive and serrated archetypes likely reflect tumorigenesis along the well-established "conventional" pathway, initiated by *APC* mutation, and the more recently described "serrated" pathway, initiated by *BRAF* activating mutations, respectively (Fearon and Vogelstein, 1990; Leggett and Whitehall, 2010). The secretory archetype may represent tumorigenesis along the less-described "alternate pathway" characterized by frequent *KRAS* activating mutations (Leggett and Whitehall, 2010). In agreement, *KRAS* mutation can promote intestinal hyperplasia and goblet cell pool expansion (Feng et al., 2011) similarly to our observations.

Our tumor archetypes bear resemblance to the CMSs proposed by the CRCSC with the notable exception that we identified an additional subtype of CIN tumors, named dARE, which is likely metabiome dependent. We hypothesize that bacterial infection by certain bacteria can induce the dARE phenotype by activating intestinal DCs to induce immune tolerance, possibly mediated by IL-10 and ARE^{+} transcript downregulation. Given that host anti-tumor immune responses greatly impact cancer dissemination and patient prognosis, further studies should consolidate our model and establish how immune modulation by the microbiota influences CRC development and treatment in patients with dARE tumors, and possibly in other cancers. It should be stressed that the identification of the dARE archetype also proved essential to our prognostication strategy: we were unable to validate CIN subtype-specific prognostic biomarkers if the transcriptionally extreme dARE tumors were not separated from the CIN tumors (data not shown).

Our dual class discovery approach also helped illustrate that the TME, rather than the CC archetype, is a major determinant of patient prognosis both between and within tumor archetypes. In particular, the aggressive stroma archetype is characterized by high transforming growth factor β (TGF-β) expression, inhibition of cytotoxic CD4/CD8 T cell activation (Thomas and Massagué, 2005), and activation of fibroblasts into cancer-associated fibroblasts (CAFs) that augments EMT processes (Calon et al., 2012). Instead, the good-prognosis SSC tumors are characterized by an active, anti-tumor immune response mediated by active cytotoxic CD4 and CD8 T cells (Figure 3). Notably, the biological processes that contribute most negatively to archetype-specific prognosis, namely, stroma activation (i.e., EMT) and reduction of immune processes (i.e., IFN-γ signaling), are also prognostic determinants within tumor archetypes. Yet, we find that the specific gene expression biomarkers that best reflect this are archetype specific rather than universal. This observation may help explain the difficulties of validating CRC biomarkers without molecular subtyping in the past. In this regard, differences in archetype distributions between public CRC cohorts may, at least in some cases, reflect differences in biopsy processing requirements: this is most clearly reflected in the laser microdissected cohorts in which, e.g., stroma archetype tumors are not identified. This calls for standardization of sample collection when the aim is subtyping and application of TME-based biomarkers, a concern recently shared by others (Dunne et al., 2016). Furthermore, only few publicly available CRC cohorts are large enough to support the study of archetype-specific biomarkers. Consequently, patient numbers in archetype-specific analysis of RFS, as presented here, are currently limited as cohorts must be divided into archetypes prior to biomarker discovery/validation. Thus, future clinical validation in prospective cohorts calls for large collaborative efforts.

Based on our results, we envision that our prognostication framework may potentially supplement TNM classification to enhance clinical patient prognostication and guide treatment decisions, e.g., to stratify patients for chemotherapeutic treatment (Figures 6D, 6E, and S6E). Today, only TNM stage III patients are routinely offered adjuvant chemotherapeutic treatment, which results in frequent undertreatment of TNM stage II patients (~25% disease relapse) and overtreatment of stage III patients (~50% disease relapse; Marshall, 2010; Puppa et al., 2010; Tsikitis et al., 2014). We foresee that TNM stage II–III patients may alternatively be stratified according to their molecular P-score, thereby enabling treatment to be directed to those patients with highest risk of relapsing. In the CRC samples analyzed here, stratification of TNM stage II patients for adjuvant treatment would dramatically help to reduce patient undertreatment: ~69% (61/89) of TNM stage II patients with relapse/CRC-death were P-score positive, while <10% (28/298) of P-score negative patients experienced relapse/CRC-death (negative predictive value [NPV] of ~91%; Figure 6E). We foresee that the P-score may similarly help reduce the current overtreatment of stage III patients given its high NPV in stage III tumors (Figure 6E). While this is promising, prospective clinical studies are needed to document the clinical benefits of the strategy. Finally, we envision that the molecular subtyping framework presented here is equally applicable in other cancer types and can also be used for development of archetype-specific treatment-predictive biomarkers.

EXPERIMENTAL PROCEDURES

Additional experimental procedures are available in the Supplemental Experimental Procedures.

Patients and Tumor Material

A total of 33 adenoma, 281 carcinoma samples, and 301 normal mucosa samples (SYSCOL cohort) from a total of 301 patients were selected from the colorectal cancer biobank at the Department of Molecular Medicine, Aarhus University Hospital, Skejby, Denmark (see Table S1 for an overview of basic molecular and clinical features for patients in the SYSCOL cohort). Patients gave their written informed consent and were followed according to national clinical guidelines. This study was conducted in accordance with local law and is approved by local institutional review boards and ethical committees.

RNA-Seq and DNA Methylation Profiling

RNA purification and sequencing of the SYSCOL cohort were performed as described (Ongen et al., 2014). Transcriptome quantification was performed using by mapping sequencing reads to human genome issue HG19 (hg19) using Tophat2 (Kim et al., 2013) and estimating fragments per kilobase of exon per million fragments mapped (FPKM) values for individual Ensembl Genes using Cufflink (Gencode v.15 annotation w/o Pseudogenes; Trapnell et al., 2010). Bacterial read counts were obtained by mapping unmapped reads (to hg19) to bacterial and viral genomes available at NCBI genome database.

DNA Methylation Profiling

DNA methylation profiles of the SYSCOL samples were generated using the infinium HumanMethylation450 BeadChip technology as previously described (Lopez-Serra et al., 2014). Raw data were processed into β values using the ChAMP R-package (Morris et al., 2014).

NMF-Based Consensus Clustering and Comparison of Predictions

NMF-based consensus clustering was performed using the R-package "NMF" (Gaujoux and Seoighe, 2010). The class number was set based on evaluation of the cophenetic coefficient, and consensus silhouette scores and samples with silhouette scores <0 were labeled as "unclassified." For identification of CC archetypes, only transcripts (HUGO Gene Nomenclature Committee [HGNC] symbols) that had a "fraction of reads of murine origin" <0.01 were analyzed (as devised by Isella et al., 2015), whereas all transcripts were included during tumor archetype discovery. The similarity of the CC and tumor archetypes, independently defined in the SYSCOL, GSE39582, and GSE41258 cohorts, was assessed using PCA of the archetype-specific area under the curve (AUC) values obtained for all common transcripts using the R-package ROCR (Sing et al., 2005) and the SubMap algorithm (Hoshida et al., 2007) via the GenePattern interface (Reich et al., 2006) using 1,000 markers genes, each null distribution, and Benjamini-Hochberg (BH)-corrected false discovery rates (FDRs). NMF-based clustering of DNA methylome data was performed using a similar strategy based on β values from the 17,711 most variable probes (interquartile range >1).

Stromal and ARE Transcripts

The ARE content of RNA transcripts were defined by the AREsite database (Gruber et al., 2011). Stromal transcripts were defined as those with a fraction of reads of murine origin >0.80 as devised by Isella et al. (2015).

Mutation and Copy Number Analysis

KRAS codon 12, 13, and 61 mutations and the mutation score were evaluated from the SYSCOL RNA-seqdata using the MuTect package (Cibulskis et al., 2013). *BRAF* exon 15 (V600E) mutations were identified by singleplex PCRs using LightScanner Master Mix and LightScanner analysis (Idaho Technology). Copy number alterations (CNAs) were estimated from the intensity of the Infinium HumanMethylation450 BeadChip by comparing adenoma-carcinoma samples to normal mucosa samples using the ChAMP R-package (Feber et al., 2014). The MSI status of the SYSCOL samples was evaluated using a Pentaplex PCR with five quasimonomorphic mononucleotide repeats and samples with three or more of five positive markers classified as MSI, as recommended in Suraweera et al. (2002).

Nearest Template Prediction

Nearest template prediction (NTP) was performed using the NTP module (Hoshida, 2010) of the GenePattern analysis toolkit (Reich et al., 2006) using default settings and a BH corrected FDR <0.25. The utilized datasets are described in the Supplemental Experimental Procedures.

Estimation of Sample Cellular Content

The cellular content of CRC samples was estimated from RNA expression and DNA methylation data using several bioinformatics methods (details in Supplemental Experimental Procedures). The following packages were used. ESTIMATE (Yoshihara et al., 2013) estimated the stroma score and ImmuneScore. CellMix (Gaujoux and Seoighe, 2013) estimated the proportion of immune cell types in CRC samples (using the basis matrix from Abbas et al., 2009). CIBERSORT (Newman et al., 2015) was used to estimate the content of leukocyte cells using the leukocyte signature matrix (LM22). The eFORGE tool v.1.1 (Breeze et al., 2016) was used to estimate the cell content of the tumor samples based on their DNA methylation profile.

Gene-Set Enrichment Analysis

GSEA was performed using GSEA v.2.2.2 provided by the Broad Institute (Mootha et al., 2003; Subramanian et al., 2005) using default settings (except minimum gene-set size = 10), and gene-set permutation type and results were visualized using EnrichmentMap (Merico et al., 2010) via the Cytoscape software v.3.2.0 (Shannon et al., 2003). Refer to Supplemental Experimental Procedures for a description of the included gene sets.

External CRC Datasets

Transcriptome data from the two external datasets used for archetype discovery (GSE39582 and GSE41258) were acquired as series matrix files from the Gene Expression Omnibus (GEO; Edgar et al., 2002; https://www.ncbi.nlm.nih.gov/geo/). The Cancer Genome Atlas (TCGA) cohorts "COAD," "READ," "BLAD," BRAD," "LUAD," and "PRAD" were acquired as level 3 (processed) data matrixes for RNA-seqV2 data from the TCGA portal (https://tcga-data.nci.nih.gov/docs/publications/tcga?). The transcriptome dataset of Burkholderia pseudomallei-infected DCs (Khaenam et al., 2014) was acquired as a series matrix file from GEO: GSE49753; https://www.ncbi.nlm.nih.gov/geo). External cohorts used for validation of prognostic biomarker panels, GSE37892, GSE14333, and GSE13294, were acquired from the CRCSC synapse data portal as frozen robust multiarray analysis (fRMA) normalized data (Guinney et al., 2015; https://www.synapse.org [Synapse id:syn2623706]), whereas GSE17538 and GSE41258 (not available at CRCSC) were acquired as series matrix files from GEO (https://www.ncbi.nlm.nih.gov/geo). See Supplemental Experimental Procedures for accession numbers for samples included in the archetype discovery and RFS analyses.

CRC Classifiers and Classification

The development and application of the CRCclassifier as well as the application of the CMSclassifier provided by the CRCSC (Guinney et al., 2015) is described in the Supplemental Experimental Procedures.

Establishment of Archetype-Specific Biomarker Panels

Details of the development and application of archetype-specific biomarker panels are described in the Supplemental Experimental Procedures.

Statistical Analysis

Unless otherwise noted, statistical significance of data was determined using a non-parametric WRS test, and $p < 0.05$ was considered significant. GSEA estimated the statistical significance by a permutation test by creating a random gene set: gene sets with FDR q values below 0.25 are generally considered significant (Mootha et al., 2003; Subramanian et al., 2005); however, a FDR limit of 0.1 is used here unless indicated in the text. For NTP analysis, samples were assigned to a template if FDR was <0.25 (BH corrected). RFS analysis was performed in TNM stage II–III patients only. RFS was measured as the interval between surgery and first recurrence or death as a result of CRC and was censored at the last follow-up or non-CRC-related death. The survival analysis and Kaplan-Meier plots were generated using Stata/IC 12.1 (StataCorp). The presented p values were evaluated by a log-rank test of equality, whereas HR and HR (95%CI) were evaluated using a Cox proportional hazards model. Multivariate Cox regression analysis was performed using tumor TNM stage, CMS type, MSI status, tumor archetype, and the relevant archetype-specific biomarker P-score as co-variables, as indicated in the text.

ACCESSION NUMBERS

The accession number for the RNA-seq data from the 314 SYSCOL adenoma and carcinoma samples reported in this paper is EGA: EGAS00001002376 (https://www.ebi.ac.uk/ega/), which is hosted by the EBI and the CRG, for controlled accesses.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.04.045.

AUTHOR CONTRIBUTIONS

J.B.B., M.E., E.T.D., T.F.Ø., and C.L.A. designed the experiments. J.B.B., M.H.R., H.O., T.B.M., M.-B.W.Ø., S.S.A., J.S., B.Ø., M.R.M., and S.L. performed the experiments and included patients. J.B.B., H.O., T.L., S.V., P.L., T.F.Ø., and C.L.A. analyzed and interpreted the data. J.B.B., T.F.Ø., and C.L.A. drafted the manuscript. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

This research is supported by grants from the European Commission *FP7* project SYSCOL (UE7-SYSCOL-258236), the Novo Nordisk Foundation (NNF16OC0023182), the Danish National Advanced Technology Foundation (056-2010-1), the John and Birthe Meyer Foundation, the Danish Council for Independent Research (Medical Sciences) (0602-02128B, DFF – 4183-00619), the Danish Council for Strategic Research (1309-00006B), and the Danish Cancer Society (R40-A1965_11_S2, R56-A3110-12-S2, R107-A7035, R133-A8520). The Danish Cancer Biobank is acknowledged for biological material. We thank P. Celis, L. Nielsen, L. Kjeldsen, B. Devantie, B. Trolle, S. Moran, D. Garcia, and C. Arribas for their technical support. The results published here are in part based upon data generated by the TCGA Research Network: https://cancergenome.nih.gov/.

Received: June 23, 2016
Revised: December 28, 2016
Accepted: April 16, 2017
Published: May 9, 2017

REFERENCES

Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H.F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE *4*, e6098.

Anastassiou, D., Rumjantseva, V., Cheng, W., Huang, J., Canoll, P.D., Yamashiro, D.J., and Kandel, J.J. (2011). Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. BMC Cancer *11*, 529.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M.L., Snajder, R., Krogsdam, A.M., Waldner, M.J., Bindea, G., Mlecnik, B., Galon, J., and Trajanoski, Z. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol. *16*, 64.

Breeze, C.E., Paul, D.S., van Dongen, J., Butcher, L.M., Ambrose, J.C., Barrett, J.E., Lowe, R., Rakyan, V.K., Iotchkova, V., Frontini, M., et al. (2016). eFORGE: A tool for identifying cell type-specific signal in epigenomic data. Cell Rep. *17*, 2137–2150.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D.V., Iglesias, M., Céspedes, M.V., Sevillano, M., Nadal, C., Jung, P., Zhang, X.H., et al. (2012). Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. Cancer Cell *22*, 571–584.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Doz, E., Lombard, R., Carreras, F., Buzoni-Gatel, D., and Winter, N. (2013). Mycobacteria-infected dendritic cells attract neutrophils that produce IL-10 and specifically shut down Th17 CD4 T cells through their IL-10 receptor. J. Immunol. *191*, 3818–3826.

Dunne, P.D., McArt, D.G., Bradley, C.A., O'Reilly, P.G., Barrett, H.L., Cummins, R., O'Grady, T., Arthur, K., Loughrey, M.B., Allen, W.L., 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, Published online May 5, 2016.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759–767.

Feber, A., Guilhamon, P., Lechner, M., Fenton, T., Wilson, G.A., Thirlwell, C., Morris, T.J., Flanagan, A.M., Teschendorff, A.E., Kelly, J.D., and Beck, S. (2014). Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. *15*, R30.

Feng, Y., Bommer, G.T., Zhao, J., Green, M., Sands, E., Zhai, Y., Brown, K., Burberry, A., Cho, K.R., and Fearon, E.R. (2011). Mutant KRAS promotes hyperplasia and alters differentiation in the colon epithelium but does not expand the presumptive stem cell pool. Gastroenterology *141*, 1003–1013.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics *11*, 367.

Gaujoux, R., and Seoighe, C. (2013). CellMix: A comprehensive toolbox for gene expression deconvolution. Bioinformatics *29*, 2211–2212.

Grade, M., Hörmann, P., Becker, S., Hummon, A.B., Wangsa, D., Varma, S., Simon, R., Liersch, T., Becker, H., Difilippantonio, M.J., et al. (2007). Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. Cancer Res. *67*, 41–56.

Gruber, A.R., Fallmann, J., Kratochvill, F., Kovarik, P., and Hofacker, I.L. (2011). AREsite: A database for the comprehensive investigation of AU-rich elements. Nucleic Acids Res. *39*, D66–D69.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356.

Harris, A.L. (2002). Hypoxia—a key regulatory factor in tumour growth. Nat. Rev. Cancer *2*, 38–47.

Hoshida, Y. (2010). Nearest template prediction: A single-sample-based flexible class prediction with confidence assessment. PLoS ONE *5*, e15543.

Hoshida, Y., Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2007). Subclass mapping: Identifying common subtypes in independent disease data sets. PLoS ONE *2*, e1195.

Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., et al. (2015). Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. *47*, 312–319.

Khaenam, P., Rinchai, D., Altman, M.C., Chiche, L., Buddhisa, S., Kewcharoenwong, C., Suwannasaen, D., Mason, M., Whalen, E., Presnell, S., et al. (2014). A transcriptomic reporter assay employing neutrophils to measure immunogenic activity of septic patients' plasma. J. Transl. Med. *12*, 65.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kishore, R., Tebo, J.M., Kolosov, M., and Hamilton, T.A. (1999). Cutting edge: Clustered AU-rich elements are the target of IL-10-mediated mRNA destabilization in mouse macrophages. J. Immunol. *162*, 2457–2461.

Laiho, P., Kokko, A., Vanharanta, S., Salovaara, R., Sammalkorpi, H., Järvinen, H., Mecklin, J.P., Karttunen, T.J., Tuppurainen, K., Davalos, V., et al. (2007). Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Oncogene *26*, 312–320.

Leggett, B., and Whitehall, V. (2010). Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology *138*, 2088–2100.

Lopez-Serra, P., Marcilla, M., Villanueva, A., Ramos-Fernandez, A., Palau, A., Leal, L., Wahi, J.E., Setien-Baranda, F., Szczesna, K., Moutinho, C., et al. (2014). A DERL3-associated defect in the degradation of SLC2A1 mediates the Warburg effect. Nat. Commun. *5*, 3608.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M., et al. (2013). Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. PLoS Med. *10*, e1001453.

Marshall, J.L. (2010). Risk assessment in Stage II colorectal cancer. Oncology (Williston Park) *24* (*1, Suppl 1*), 9–13.

Mårtensson, A., Oberg, A., Jung, A., Cederquist, K., Stenling, R., and Palmqvist, R. (2007). Beta-catenin expression in relation to genetic instability and prognosis in colorectal cancer. Oncol. Rep. *17*, 447–452.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. PLoS ONE *5*, e13984.

Moffitt, R.A., Marayati, R., Flate, E.L., Volmar, K.E., Loeza, S.G., Hoadley, K.A., Rashid, N.U., Williams, L.A., Eaton, S.C., Chung, A.H., et al. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat. Genet. *47*, 1168–1178.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., and Beck, S. (2014). ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics *30*, 428–430.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., et al. (2014). Putative cis-regulatory drivers in colorectal cancer. Nature *512*, 87–90.

Puppa, G., Sonzogni, A., Colombari, R., and Pelosi, G. (2010). TNM staging system of colorectal carcinoma: A critical appraisal of challenging issues. Arch. Pathol. Lab. Med. *134*, 837–852.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Rigby, R.J., Knight, S.C., Kamm, M.A., and Stagg, A.J. (2005). Production of interleukin (IL)-10 and IL-12 by murine colonic dendritic cells in response to microbial stimuli. Clin. Exp. Immunol. *139*, 245–256.

Saridaki, Z., Souglakos, J., and Georgoulias, V. (2014). Prognostic and predictive significance of MSI in stages II/III colon cancer. World J. Gastroenterol. *20*, 6809–6814.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Sheffer, M., Bacolod, M.D., Zuk, O., Giardina, S.F., Pincas, H., Barany, F., Paty, P.B., Gerald, W.L., Notterman, D.A., and Domany, E. (2009). Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. Proc. Natl. Acad. Sci. USA *106*, 7131–7136.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Suraweera, N., Duval, A., Reperant, M., Vaury, C., Furlan, D., Leroy, K., Seruca, R., Iacopetta, B., and Hamelin, R. (2002). Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. Gastroenterology *123*, 1804–1811.

Thomas, D.A., and Massagué, J. (2005). TGF-beta directly targets cytotoxic T cell functions during tumor evasion of immune surveillance. Cancer Cell *8*, 369–380.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Tsikitis, V.L., Larson, D.W., Huebner, M., Lohse, C.M., and Thompson, P.A. (2014). Predictors of recurrence free survival for patients with stage II and III colon cancer. BMC Cancer *14*, 336.

Yang, Q., Bermingham, N.A., Finegold, M.J., and Zoghbi, H.Y. (2001). Requirement of Math1 for secretory cell lineage commitment in the mouse intestine. Science *294*, 2155–2158.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612.