Redefining tumor classification and clinical stratification through a colorectal cancer single-cell atlas

Ateeq M. Khaliq[1,*], Zeyneb Kurt[2,*], Miles W. Grunvald[1,*], Cihat Erdogan[3], Sevgi S. Turgut[3], Tim Rand[4], Sonal Khare[4], Jefferey A. Borgia[1], Dana M. Hayden[1], Sam G. Pappas[1], Henry R. Govekar[1], Anuradha R. Bhama[1], Ajaypal Singh[1], Richard A. Jacobson[1], Audrey E. Kam[1], Andrew Zloza[1], Jochen Reiser[1], Daniel V. Catenacci[5], Kiran Turaga[5], Milan Radovich[6], Sheeno Thyparambil[7], Mia A. Levy[1], Janakiraman Subramanian[8], Timothy M. Kuzel[1], Anguraj Sadanandam[9], Arif Hussain[10], Bassel El-Rayes[11], Ameen A. Salahudeen[4]✉, Ashiq Masood[1]✉

Affiliations:

[1]Rush University Medical Center, Chicago, IL, USA.

[2]Northumbria University, Newcastle Upon Tyne, UK.

[3]Isparta University of Applied Sciences, Isparta, Turkey.

[4]Tempus Labs, Inc., Chicago, IL, USA.

[5]The University of Chicago, Chicago, IL, USA.

[6]Indiana University School of Medicine, Indianapolis, IN, USA.

[7]mProbe Inc. Rockville, Maryland, USA

[8]University of Missouri-Kansas City School of Medicine, Kansas City, MO, USA.

[9]Institute of Cancer Research, London, UK.

[10]University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center, Baltimore, MD USA.

[11]Emory University Winship Cancer Institute, Atlanta, GA, USA.

[*]These authors contributed equally: Ateeq M. Khaliq, Zeyneb Kurt and Miles W. Grunvald.

✉e-mail: ashiq_masood@rush.edu; ameen@tempus.com

ABSTRACT

Colorectal cancer (CRC), a disease of high incidence and mortality, exhibits a large degree of inter- and intra-tumoral heterogeneity. The cellular etiology of this heterogeneity is poorly understood. Here, we generated and analyzed a single-cell transcriptome atlas of 49,859 CRC cells from 16 patients, validated with an additional 31,383 cells from an independent CRC patient cohort. We describe subclonal transcriptomic heterogeneity of CRC tumor epithelial cells, as well as discrete stromal populations of cancer-associated fibroblasts (CAFs). Within CRC CAFs, we identify the transcriptional signature of specific subtypes (CAF-S1 and CAF-S4) that significantly stratifies overall survival in more than 1,500 CRC patients with bulk transcriptomic data. We also uncovered two CAF-S1 subpopulations, *ecm-myCAF* and *TGFß-myCAF*, known to be associated with primary resistance to immunotherapies. We demonstrate that scRNA analysis of malignant, stromal, and immune cells exhibit a more complex picture than portrayed by bulk transcriptomic-based Consensus Molecular Subtypes (CMS) classification. By demonstrating an abundant degree of heterogeneity amongst these cell types, our work shows that CRC is best represented in a transcriptomic continuum crossing traditional classification systems boundaries. Overall, this CRC cell map provides a framework to re-evaluate CRC tumor biology with implications for clinical trial design and therapeutic development.

66 **Main**

67 Colorectal cancer (CRC) is the third most commonly diagnosed cancer and a leading cause of
68 cancer-related mortality worldwide[1,2]. Approximately 50% of patients experience disease relapse
69 following curative-intent surgical resection and chemotherapy[3,4]. Despite the high incidence and
70 mortality of advanced CRC, few effective therapies have been approved in the past several
71 decades[5]. One barrier to the development of efficacious therapeutics is the biological
72 heterogeneity of CRC and its variable clinical course[6]. While landmark studies from The Cancer
73 Genome Atlas (TCGA) have defined the somatic mutational landscape within CRC, several
74 studies have shown that stromal signatures, including fibroblasts and cytotoxic T cells, are likely
75 the main drivers of clinical outcomes[7–12]. These findings suggest that the clinical phenotypes of
76 CRC and by extension, its tumor biology, is shaped by a complex niche of heterotypic cell
77 interactions within the tumor microenvironment (TME)[8-12].

78

79 Bulk gene expression analyses by several independent groups have identified distinct CRC
80 subtypes[13–15]. Reflecting both the tumor and TME, an international consortium published the
81 Consensus Molecular Subtypes (CMS), which proposed four distinct subtypes of CRC[15].
82 Unfortunately, the association between CMS and meaningful therapeutic response to specific
83 agents have shown inconsistent results across studies and CMS lacks a concordance between
84 primary and metastatic CRC tumors, limiting its utility thus far in clinical decision making[16–23]. As
85 a result, an improved CMS classification or an alternative classification system is required to
86 improve clinical utility.

87

88 To overcome the limitations of bulk-RNA sequence profiling, we utilized single-cell RNA
89 sequencing (scRNA-seq) to more thoroughly evaluate the CRC subtypes at the molecular level,
90 including within the context of the currently defined CMS classification. We dissected heterotropic
91 cell states of tumor epithelia and stromal cells, including a cancer-associated fibroblast

92 (CAF) population. The CAF population's clinical and prognostic significance became apparent
93 when CAF signatures were applied to large, independent CRC transcriptomic cohorts.
94
95 **RESULTS**
96 We profiled sixteen primary colon tissue samples and eight adjacent non-malignant tissues (24 in
97 total) using droplet-based, scRNA-seq. Altogether, we captured and retained 49,589 single cells
98 after performing quality control for downstream analysis (**Fig. 1a, supplementary table 1**). All
99 scRNA-seq data were merged and normalized to identify robust discrete clusters of epithelial cells
100 (*EPCAM+, KRT8+,* and *KRT18+*), fibroblasts (*COL1A2+*), endothelial cells (*CD31+*), T cells
101 (*CD3D+),* B cells *(CD79A+),* and myeloid cells (*LYZ+*) using canonical marker genes. Additionally,
102 each cell type compartment was analyzed separately. Cluster v0.4.1 and manual review of
103 differentially expressed genes in each subcluster were studied to choose the best cluster
104 resolution without cluster destabilization (see methods)[24]. Cell population designation was
105 chosen by specific gene expression, and SingleR was also utilized for unbiased cell type
106 recognition (see methods)[25–28].
107
108 In addition to cancer cells, we identified diverse TME cell phenotypes, including fibroblasts
109 subsets (*CAF-S1* and *CAF-S4*), endothelial cells, CD4+ subsets (*naïve/memory, Th17,* and
110 *Tregs*), CD8+ subsets (*naïve/memory, cytotoxic, tissue-resident memory,* and *Mucosa-*
111 *Associated Invariant (MAIT) cells*), NK cells, innate lymphoid cell (ILC) types, B cell phenotypes
112 (*naïve, memory, germinal center,* and *plasma cells*), and monocyte lineage phenotypes (*C1DC+*
113 *dendritic cells, proinflammatory monocyte*s [*IL1B, IL6, S100A8,* and *S100A9]*), and M2 polarized
114 anti-inflammatory [*CD163, SEPP1, APOE,* and *MAF]*), tumor-associated macrophages (TAMs)
115 (**Fig. 1b-d, Extended Data Figs. 1-3, and Extended Data Tables 1-4**)[25–28].
116

117 For validation, we additionally profiled 31,383 high-quality, single cells from an independent cohort
118 using stringent criteria to corroborate our findings (see methods)[29]. Thus, a total of 81,242 high-
119 quality cells were profiled to produce a single-cell map of 40 colorectal cancer patients. The results
120 of the primary CRC cohort (49,859 single-cells) are available at the Colon Cancer Atlas
121 (www.websiteinprogress.com).
122
123 **Malignant colon cancer reveals tumor epithelial cell subclonal heterogeneity and**
124 **stochastic behavior.**
125
126 We detected 8,965 tumor and benign epithelial cells (*EPCAM+, KRT8+,* and *KRT18+*) and, on
127 reclustering, produced 17 epithelial clusters (designated C1 to C17) (**Fig. 2a and Supplementary**
128 **Table 2**). Clusters were chiefly influenced by colonic epithelial markers, including those for
129 stemness (*LGR5, ASCL2, OLFM4,* and *STMN1*), enterocytes (*FABP1* and *CA2*), goblet cells
130 (*ZG16, MUC2, SPINK4,* and *TFF3*), and enteroendocrine cells (*PYY* and *CHGA*)
131 (**Supplementary Fig. 2a and Supplementary Table 2**). Tumor cells exhibited a high degree of
132 de-differentiated state of plasticity possibly accounting for lasting cancer growth (**Supplementary**
133 **Fig. 2b**)[30]. Each distinct tumor-derived cluster was mostly patient-specific, reflecting a high degree
134 of inter-patient tumoral cell heterogeneity. In contrast, epithelial populations derived from non-
135 malignant tissue samples from multiple patients clustered together, a pattern observed in previous
136 studies confirming both normal tissue homeostasis and limited sample batch effects (**Fig. 2a**)[31,32].
137
138 We next aimed to identify gene expression programs shared across these clusters using hallmark
139 pathway analysis[33]. A strong overlap was observed for multiple pathways such as activation of
140 inflammatory, epithelial-mesenchymal transformation (EMT), immune response, and metabolic
141 pathways (**Fig. 2b**). Interestingly, high microsatellite instability (MSI-H) and microsatellite stable
142 (MSS) CRC tumors, considered clinically separate entities, demonstrated similar pathway

143 program activation within the tumor epithelial populations. Some clusters also showed activation
144 of unique pathways such as activation of apical junctions and angiogenesis (C6), hypoxia and
145 fatty acid metabolism (C11) and Notch signaling and DNA repair (C14), among others. However,
146 MSI-H tumors differed from MSS tumors based on immune cell infiltration (**Extended Data Fig.**
147 **1**).
148
149 Since intratumoral heterogeneity is recognized as a key mechanism contributing to drug
150 resistance, cancer progression, and recurrence, we next focused on dissecting potential
151 transcriptomic states to identify heterogeneity within each patient's tumor[34–37]. We found that each
152 tumor specimen contained 2-10 distinct tumor epithelial clusters (**Fig. 2c**). Gene set variation
153 analysis (GSVA) was performed on cells from individual tumor samples and illustrated the sub-
154 clonal transcriptomic heterogeneity within each specimen (**Supplementary Fig. 2c**)[38]. Clusters
155 identified in individual pathway analysis demonstrated the up- or down-regulation of crucial
156 metabolic and oncogenic pathways between samples, suggesting wide phenotype variations
157 between cells from the same tumor[39].
158
159 Given the evidence of intratumoral epithelial heterogeneity, we next performed trajectory
160 inference using pseudotime analysis to identify potential alignments or lineage relationships (i.e.,
161 right versus left-sided CRC), CMS classification, or MSI status
162 (**Fig. 2d**)[40,41]. This analysis also served as a control for inter-patient genomic heterogeneity and
163 provided an orthogonal strategy to confirm the transcriptomic trends we identified. We detected
164 five molecular states (S1n/t to S5n/t) with malignant and normal epithelial cells intermixed along
165 a joint transcriptional trajectory. This observation is consistent with prior studies demonstrating
166 that CRC cells recapitulate normal colon epithelia's multilineage differentiation process as each
167 transcriptional state's pathway activation in both normal and tumor cells was related to normal

168 colon epithelial function of nutrient absorption or maintaining colon homeostasis (**Supplementary**
169 **Fig. 3 and Supplementary Table 3**)[42,43].
170
171 Additionally, tumor cells showed upregulation of embryogenesis (S2t), consistent with previous
172 findings that tumor cells revert to their embryological states in cancer development
173 (**Supplementary Table 4**)[44]. Interestingly, there were no significant associations with anatomic
174 location, CMS classification, or MSI status within our dataset or an independent dataset of 31,383
175 single cells (**Supplementary Fig. 4**)[29]. Hence, in our analysis, CRC development mainly
176 represents a hijacking of the normal epithelial differentiation program, coupled with the acquisition
177 of embryogenic pathways (**Supplementary Fig. 3**)[45].
178
179 **CRC-associated fibroblasts in the tumor microenvironment exhibit diverse phenotypes,**
180 **and specific subtypes are associated with poor prognosis.**
181
182 We next focused on CRC TME subpopulations. High-quality 819 fibroblasts were re-clustered into
183 eight clusters, and then phenotypically classified into two major subtypes to assess for further
184 CAF heterogeneity. These phenotypic subtypes were found to be immunomodulatory CAF-S1
185 (*PDGFRA*+ and *PDPN*+) and contractile CAF-S4 (*RGS5*+ and *MCAM*+) (**Fig. 3 and**
186 **Supplementary Table 4**)[46]. This fibroblast cluster dichotomy was also observed in the
187 independent CRC patient scRNA-seq dataset of 31,383 cells (**Supplementary Fig. 5**)[29].
188
189 The CAF-S1 and CAF-S4 subtypes showed striking resemblances to the mCAF (extracellular
190 matrix) and vCAF (vascular) fibroblast subtypes, respectively, as previously described in a mouse
191 breast cancer model[47]. Most clusters were found in multiple patients, albeit in varying proportions,
192 signifying shared patterns in CAF transcriptomic programs between patients. Fibroblasts derived
193 from MSI-H tumors were distributed similarly throughout these clusters.

194 CAF-S1 exhibited high chemokine expressions such as *CXCL1, CXCL2, CXCL12, CXCL14,* and
195 immunomodulatory molecules including *TNFRSF12A* (**Supplementary Fig. 6**). Additionally,
196 CAF-S1 expressed extracellular matrix genes including matrix-modifying enzymes (*LOXL1* and
197 *LOX*)[47]. To determine this population's functional significance, we compared the CAF-S1
198 population transcriptomes to those described recently in breast cancer, lung cancer, and head
199 and neck cancer[48]. We recovered five CAF subtypes, within the CAF-S1 population, including
200 *ecm-myCAF* (extracellular; *GJB*+), *IL-iCAF* (growth factor, *TN*F and interleukin pathway;
201 *SCARA5*+), *detox-iCAF* (detoxification and inflammation; *ADH1B*+), *wound-myCAF* (collagen
202 fibrils and wound healing; *SEMA3C*), and *TGFβ-myCAF* (*TGF-β* signaling and matrisome;
203 *CST1+, TGFb1+*), which were previously divided into two major subtypes: iCAF and *myCAF* (**Fig.**
204 **3b**). Among these five subtypes, *ecm-myCAF* and *TGFβ-myCAF* are known to correlate with
205 immunosuppressive environments and are enriched in tumors with high regulatory T lymphocytes
206 (Tregs) and depleted CD8+ lymphocytes. Additionally, these subtypes are associated with
207 primary immunotherapy resistance in melanoma and lung cancer[48].
208
209 The CAF-S4 population expressed pericyte markers (*RGS5+, CSPG4+,* and *PDGFRA*+), *CD248*
210 *(endosialin)*, and *EPAS1 (HIF2-α),* that this particular CAF subtype is vessel-associated, with
211 hypoxia potentially contributing to invasion and metastasis, as has been shown in another study
212 [47]. CAF-S4 clustered into the immature phenotype (*RGS5+, PDGFRB+,* and *CD36+)* and the
213 differentiated myogenic subtype (*TAGLN*+ and *MYH11*+)(**Fig. 3c and Supplementary Table**
214 **4**)[49].
215
216 Given the correlation between CMS4 and fibroblast infiltration, we next sought to test the
217 existence of CAF-S1 and CAF-S4 signatures in bulk transcriptomic data and their association with
218 clinical outcomes[15]. To this end, we interrogated and carried out a meta-analysis of eight
219 colorectal cancer transcriptomic datasets comprising 1,584 samples. We detected a strong and

220 positive correlation between specific gene expressions characterizing each CAF subtype in CRC.
221 We also confirmed the presence of CAF-S1 and CAF-S4 signatures in pancreatic
222 adenocarcinoma (n=118) and non-small cell lung cancer (NSCLC, n=80) cohorts (**Fig. 4**) (see
223 methods for datasets). The gene signatures were specific to each CAF-S1 and CAF-S4, thus
224 confirming their existence in TME of CRC and other tumor types.
225
226 We found high CAF-S1 and CAF-S4 signatures associated with poor median overall-survival
227 (HR>1, $p<0.05$), irrespective of CMS subtypes in three independent CRC datasets
228 (**Supplementary Fig 7**). Additionally, CAF signatures stratified the CMS4 subtype into high- and
229 low-risk overall survival in all datasets, thus identifying additional heterogeneity and providing
230 prognostication in this aggressive patient subgroup (**Fig. 4b-d**). Here, using scRNA-seq, we show
231 for the first time that high CAF infiltration in CRC is associated with poor prognosis across all
232 molecular subtypes, and which further stratifies the CMS4 subgroup into high and low-risk clinical
233 phenotypes in CRC cohorts.
234
235 **Single-cell RNA sequencing reveals heterogeneity beyond Consensus Molecular Subtypes**
236 **in colorectal cancers and offers therapeutic opportunities.**
237
238 The lack of association between tumor epithelia and CMS classification, as well as the survival
239 differences between high- and low-risk CAF signatures across CRC molecular subtypes suggest
240 CRCs are much more heterogeneous than the traditional classification systems have indicated
241 (e.g. those systems defined by somatic alterations, epigenomic features, and bulk gene
242 expression data)[13,14,50,51].
243
244 Among these the widely adopted CMS classification, which reflects both the malignant cell
245 phenotypes and the TME, classified CRC into CMS1 (MSI immune), CMS2 (canonical), CMS3

246 (metabolic), and CMS4 (mesenchymal) subtypes based on bulk transcriptomic signatures[15]. To
247 test our hypothesis, we estimated every cell type fraction using single-cell data from eight pooled
248 datasets (>1,500 samples) with a machine-learning algorithm, CIBERSORTx[52]. When we
249 compared epithelial, immune, and stromal cell populations among the CMS subtypes, we did not
250 detect a distinct pattern of tumor, immune, or stromal cell signatures across the different CMS
251 subtypes. Each CMS subtype was enriched in these cell types in varying proportions but without
252 a clear distinction between the four subtypes, suggesting a lack of clear separation among the
253 CMS subsets at the single-cell resolution (**Fig. 5a-b**). Upon analysis of four independent bulk RNA
254 datasets, there was significant discordance in terms of cell phenotype enrichment with respect
255 to each CMS subtype across the datasets except CMS4 which had predominant stromal
256 enrichment (**Supplementary Figs. 8-11)**[53]. These discordant results could be due to intra-patient
257 CMS heterogeneity, intratumoral variation in tumor purity, stromal and immune cell infiltration,
258 and CMS's inability to address tumor/TME-to-tumor/TME variability, among others[54]. Thus, novel
259 approaches that consider these factors are required to stratify patients for optimal biomarker and
260 therapeutic development.
261
262 Based on the above findings, we postulate that CRC is more accurately represented in a
263 transcriptomic continuum previously proposed by Ma et al.[55]. The authors analyzed bulk
264 transcriptomic data using a novel computational framework in which denovo, unsupervised
265 clustering methods (k-medoid, non-negative-matrix factorization, and consensus clustering)
266 demonstrated the existence of CRC in a transcriptomic continuum[56–58]. They further carried out
267 principal component analysis and robustly validated two principal components, PC Cluster
268 Subtype Scores 1 and 2 (PCSS1 and PCSS2, respectively). We reasoned that using single-cell
269 data could deepen our understanding of how each cell phenotype contributes to the CRC tumor
270 microenvironment using continuous scores that inform CRC diversity beyond binning CRC into
271 traditional classifications.

272

273 We evaluated every cell fraction (epithelial, stromal, and immune components) from our data
274 with the Ma et al. algorithm on eight pooled bulk transcriptomic datasets, focusing on PCSS1
275 and PCSS2, since these were validated in the original study (**Fig. 5c-d**). On projecting single-
276 cell expression profiles on quadrants corresponding to each of the four CMS, we noted a lack of
277 separation of all cell phenotypes between CMS subtypes suggesting that CRC exists in a
278 transcriptomic continuum[55].

279

280 To test continuous scores reproducibility, we analyzed four bulk transcriptomic datasets
281 separately; we found that transcriptional shifts were reproducible across datasets for all major cell
282 types (**Supplementary Figs. 8-11)**. The continuous scores showed no reliability in classifying
283 CRC into immune–stromal rich (CMS1/CMS4) or immune-stromal desert (CMS2/CMS3) subtypes
284 as proposed previously [20]. Thus, confirming continuous scores rather than discrete subtypes may
285 improve classifying CRC tumors and may explain tumor-to-tumor variability, tumor/TME-to-
286 tumor/TME and TME-to-TME variabilities [55]. Of note, CAF-S1 exhibited high PCSS1 and PCSS2
287 scores across independent datasets, correlating with the CMS4 subtype. Thus, our analysis
288 identified CAF-S1 as a cell of origin for biological heterogeneity in CMS4 subtypes associated
289 with poor prognosis (**Supplementary table 5**).

290

291 **DISCUSSION**

292 In the present study, we evaluated the CMS classification of CRC that have been developed by
293 bulk RNA-seq through single-cell resolution transcriptomic analysis. We find that stromal cells
294 engender a more significant contribution to biological heterogeneity. Although previous studies
295 employing bulk transcriptomics have demonstrated that the degree of stromal infiltration is
296 associated with prognosis and a small scRNAseq study utilizing 26 fibroblasts demonstrated poor
297 survival among CAF-enriched CRC tumors especially the CMS4 subtype. Here, using much larger

298 sample set of 1,182 high-quality fibroblasts, we identify the CAF-S1 subtype to be the cell of origin
299 associated with poor prognosis across all CRC CMS subtypes and not just CMS4 [59]. Further, we
300 developed a novel signature of CAF infiltration and demonstrate that CMS4 can be stratified into
301 risk groups associated with good or poor median overall survival. These findings are significant
302 since the CMS4 subtype, is primarily stromal-driven and is enriched in more than 40% of
303 metastatic CRC samples from patients with worse outcomes[21]. We also identified CAF-S1
304 subtypes associated with certain biological functions in other cancers, including the *ecm-myCAF*
305 and *TGFβ-myCAF* subtypes (responsible for immunotherapy resistance in NSCLC and
306 melanoma), and CAF-S4s known to play a role in inducing cancer cell invasion[48,49]. Thus,
307 targeting CAFs to remodel the tumor microenvironment may lead to improved and much-needed
308 therapeutic development for metastatic CRC patients[21,22].
309
310 Targeting of CAFs in solid tumors is being explored in multiple clinical trials with variable results[60–
311 62]. Such studies likely failed to address CAF heterogeneity and their complex interactions with the
312 other cells of TME. Our study suggests that CRC may be intricately entwined with the stroma,
313 and therefore may be amenable to stromal targeted combinatoric approaches, including
314 monoclonal antibodies that abrogate CAF-S1 function. In future studies, the treatment of CRC
315 patients should involve stroma targeted therapies and take the above aspects into
316 consideration[60,63]. The scRNA-seq or bulk-RNA-seq signatures corresponding to CAF-S1 and
317 CAF-S4 may serve as suitable biomarkers for tumors that are reliant on this axis. Immunotherapy
318 responses in MSS CRC, which comprise almost 95% of metastatic CRC, are lacking; CAF
319 subpopulations within the TME may be suppressing immune responses in these tumors [64]. Based
320 on our analysis, we speculate that targeting CAF-derived chemokines and cytokines via
321 biospecific antibodies, vaccines, or even cell-based therapies, may enhance current checkpoint
322 blockade strategies[60]. Functional validation and clinical studies will be required to confirm the
323 clinical utility of targeting these CAF populations in CRC.

More importantly, our study's single-cell resolution enables us to investigate whether tumor cell transcriptomes, and by extension, biological phenotypes, are the primary determinant of CMS classification. Based on our findings, it appears that bulk analysis may have been confounded by varying tumor microenvironment population enrichment, and that tumor cells within each patient do not segregate into static phenotypes but rather exhibit considerable plasticity. In contrast, the single-cell analysis uncovered the complex and mixed cellular-phenotypes among each cell specific subpopulation, which projected in a transcriptomic continuum across CMS subtypes. These findings were further supported by scRNA-seq CMS classification analysis that assigned each CRC sample to multiple CMS subtypes thereby suggesting CMS heterogeneity in each CRC tumor[20,65]. These findings may also explain why CMS-defined populations of tumors have not been readily observed in transcriptomic data from independent CRC cohorts[21,53]. Our data indicate that attempts to divide CRC phenotypes into the current discrete subtypes may undermine optimal patient stratification in the clinical trial setting. Intriguingly, by applying two independent algorithms, we demonstrate that CRC tumors and their ecosystems exist in a transcriptomic continuum and not only show tumor-to-tumor variability (as proposed by Ma et al.) but also demonstrate tumor/TME-to-tumor/TME transcriptional variability at the single-cell resolution[55]. The continuous scores are reproducible across transcriptomic datasets, thus allowing robust identification of patient subtypes. This may help to optimize CRC treatment in future studies.

In conclusion, our study lends strong support to the tumor biology models proposed by Ma et al. (and other groups) and represents a conceptual shift in our understanding of CRC pathogenesis, clinical management, and therapeutic development. Future studies will need to consider tumor-TME to tumor-TME heterogeneity which will be critical for optimizing biomarkers and treatment strategies for CRC.

ACKNOWLEDGEMENTS

This study was supported by the startup fund provided to A.M. by the Rush University Medical Center; the OCM grant to A.M by Rush University Cancer Center. Part of A.H.'s time was supported by a Merit Review Award (I01 BX000545) from the Medical Research Service, Department of Veterans Affairs. We would also like to thank Dr. Levi Waldron for sharing code from his publication entitled, "Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis." Above all we want to thank our patients who participated in this study and their families.

AUTHOR CONTRIBUTIONS

A.M. devised, supervised the study, and wrote the manuscript. A.M.K. performed data analyses, wrote the manuscript, and created figures. Z.K. performed data analyses and wrote the manuscript. M.W.G. aided in analysis, wrote the manuscript, and generated figures. A.S. supervised study and wrote manuscript. C.E. and S.S.T. helped with bulk transcriptomic analysis. D.M.H, H.R.G., A.R.B helped with sample collection. All other authors contributed substantially to data interpretation, and manuscript editing. All authors read and approved this manuscript.

CODE AVAILABILITY

The code generated and utilized in the completion of this publication will be available in a Github repository specific to this project.

DATA AVAILABILITY

Sequencing data deposition is currently in progress. Ten bulk transcriptomic datasets were accessed from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

COMPETING INTERESTS

A.M. and J.A.B. received research funding from Tempus lab.

A.S. receives research funding from Bristol-Myers Squibb; Merck KGaA, Pierre Fabre. Further, A.S. holds patent PCT/IB2013/060416, '*Colorectal cancer classification with differential prognosis and personalized therapeutic responses*' and patent number 2011213.2 '*Prognostic and Treatment Response Predictive Method*.'

383 **METHODS**

384

385 ***Patient and tissue sample collection.*** Patients with resectable untreated CRC who underwent
386 curative colon resection at Rush University Medical Center (Chicago, IL, USA) were included in
387 this Institutional Review Board (IRB)-approved study. CRC specimens from 16 patients including
388 nine Caucasian, six African American and one Asian patient with corresponding 8 adjacent normal
389 tissue samples were processed immediately after collection at Rush University Medical Center
390 Biorepository and sent for scRNA-seq. Thus, our scRNA-seq atlas represent diverse patient
391 population. The study was conducted in accordance with ethical standards and all patients
392 provided written informed consent.

393

394 ***Droplet based scRNA-seq - 10× library preparation and sequencing.*** Single-cell RNA
395 sequencing (scRNA-seq) was performed using 10X Genomics Single Cell 5' Platform. Tumors
396 and non-malignant samples were enzymatically dissociated (*Miltenyi*), filtered through a 70-
397 micron cell strainer, pelleted after centrifugation at 300 x*g* and resuspended in DAPI-FACS buffer
398 (PBS, 0.04% BSA). Samples were sorted and viable singlets were gated on the basis of scatter
399 properties and DAPI exclusion. Approximately 3000 cells were pelleted and resuspended in PBS,
400 and cells underwent single cell droplet-based capture on 10X Chromium instruments according
401 to the 10X Genomics Single Cell 5' Platform protocol. Transcriptome libraries post-fragmentation,
402 end-repair, and A-tailing double-sided size selection, and subsequent adaptor ligation also
403 followed the manufacturer's protocol. Illumina *NextSeq 550* was used for library sequencing and
404 data were mapped and counted using Cellranger-v3.1.0 (*GRCh38/hg38*).

405

406 ***scRNA-seq data quality control, gene-expression quantification, dimensionality reduction,***
407 ***and identification of cell clusters.*** *Cell Ranger* was utilized to process the raw gene expression
408 matrices per samples and all samples from multiple patients were combined in R package (v3.6.3

409 2020-02-29] -- "*Holding the Windsock*"). Seurat package (v3.2.2) was used in this integrative
410 multimodal analysis[66]. Genes detected in fewer than three cells and cells expressing less than
411 200 detected genes were filtered out and excluded from analysis. In addition, cells expressing >
412 25% mitochondria were removed. Cell cycle scoring was performed, (for the S phase and the
413 G2M phase) and the predicted cell cycle phases were calculated. Doublet detection and any
414 higher-order multiplets that were not dissociated during sample preparation were removed via
415 the *DoubletFinder* (v2.0.2) package using default settings[67]. Following quality control one non-
416 malignant colon sample (B-cac13) was discarded due to poor data quality. Finally, 49,859 cells
417 remained and were utilized for downstream analysis.
418 We adopted the general protocol described in Stuart et al. (2019) to group single cells into different
419 cell subsets[66]. We employed the following steps: clustering the cells within each compartment
420 (including the selection of variable genes for each dataset based on a variance stabilizing
421 transformation [VST]), canonical correlation analysis (CCA) to remove batch effects among the
422 samples, reduction of dimensionality, and projection of cells onto graphs [68,69]. Principal
423 component analysis (PCA) was carried out on the scaled data of highly variable genes[70]. The first
424 30 principal components (PCs) were used to cluster the cells and to perform a subtype analysis
425 by nonlinear dimensionality reduction (t-SNE) and to construct Uniform Manifold Approximation
426 and Projection (UMAP) for cell embeddings[71,72]. We identified cell clusters under the optimal
427 resolution by a shared nearest neighbor (SNN) modularity optimization-based clustering method.
428 We implemented the *FindClusters* function of the Seurat package, which first calculated *k-nearest*
429 *neighbors* and constructed the SNN graph. We implemented the original *Louvain algorithm*
430 (algorithm = 1) for modularity optimization. Additionally, we utilized Clustree (v0.4.3) and manual
431 review for identifying the best clustering resolution[24].
432
433 ***Major cell type detection and data visualization.*** To identify all major cell types, we evaluated
434 differentially expressed markers in each identity cell group by comparing them to other clusters

435 using the Seurat *FindAllMarkers* function. We used positively expressed genes with an average
436 expression of >/= 2-fold higher in that subcluster than the average expression in the rest of the
437 other subclusters. We used known marker genes, which have the highest fold expression in that
438 cluster with respect to the other clusters. We also utilized SingleR ((v0.99.10, R Package), which
439 leverage large transcriptomic datasets of well-annotated cell types and manual annotation for
440 cell-type identification[31,73–75]. Depending on the presence of known marker genes the clusters
441 were grouped as: epithelial cells (*EPCAM, KRT8,* and *KRT18*), fibroblasts (*COL1A1, DCN,*
442 *COL1A2,* and *C1R*), endothelial cells (CD31+), myeloid cells (*LYZ, MARCO, CD68,* and
443 *FCGR3A*), CD4 T cells (*CD4*), CD8 T cells (*CD8A* and *CD8B*), and B cells (*MZB1*), [31,47,73,76–80].
444 The cells were eventually assembled into DGE matrices within each compartment, containing all
445 six cell types.
446
447 ***Major-cell type subclustering and data visualization.*** Each major cell type, including epithelial
448 cells, endothelial cells, T cells, B cells, myeloid cells, and fibroblasts was reclustered and
449 reanalyzed to study each compartment at a higher resolution to detect granular cellular
450 heterogeneity in CRC. Clustree (v0.4.3) and manual review were utilized for optimal cluster
451 detection. For cell annotation of each cell type, we utilized published literature gene expression
452 signatures and manual review of differential genes among clusters. Additionally, we again utilized
453 SingleR (v0.99.10, R Package) for unbiased cell annotation. Interestingly, reclustering of major
454 compartments individually also detected clusters expressing hybrid markers as well as cell
455 clusters expressing markers from distinct lineages (such as T cell clusters expressing B cells);
456 these were removed and excluded for further analysis. We utilized UMAP for visualization
457 purposes. For validation, we analyzed 65,362 cells from 24 patients and applied the same quality
458 control metrics as outlined above, retaining 31,383 high-quality single cells for further analysis[29].
459 These high-quality cells were analyzed utilizing the same pipelines and parameters as that for our
460 primary cohort (**Supplementary Figs. 4-5 and 12-13**).

461 The InferCNV (v1.2.1) package was used with default paramets to identify the evidence for
462 somatic large-scale chromosomal copy number alteration in epithelial cells (*EPCAM+, KRT8+,*
463 *KRT18+*)[81]. Non-malignant epithelial cells were used as the control group.
464
465 ***Trajectory analysis.*** We used Monocle v.2 (v2.14.0), a reverse graph embedding method to
466 reconstruct single-cell trajectories in tumor and non-malignant epithelium[82]. In brief, we used UMI
467 count matrices and the *negbinomial.size*() parameter to create a *CellDataSet* object in the default
468 setting. We grouped projected cells on UMAP in default settings for visualization of monocle
469 results. We defined the cumulative duration of the trajectory to show the average amount of
470 transcriptional transition that a cell undergoes as it passes from the starting state to the end state.
471 The cells were also ordered in pseudotime to explain the transition of cells from one state to
472 another.
473
474 ***Pathway- Gene set variation analysis (GSVA).*** Pathway analysis was performed on the 50
475 hallmark gene sets downloaded from *Molecular Signatures Database (v7.2)*. We used GSVA
476 (v1.34.0), a non-parametric, unsupervised method to estimate the gene set variations and
477 evaluation of pathway enrichment, and pathway scores were calculated for each cell using
478 standard settings [33,38].
479
480 ***DNA and bulk RNA library construction.*** DNA and bulk RNA sequencing was performed as
481 previously described[83]. One hundred nanograms of DNA from each tumor was mechanically
482 sheared to an average size of 200 bp. Using the *KAPA Hyper Prep Pack*, DNA libraries were
483 packed, hybridized into the *xT probe* package, and amplified with the *KAPA HiFi HotStart*
484 *ReadyMix*. For uniformity, each sample needed to have 95% of all targeted base pairs sequenced
485 to a minimum depth of 300x. One hundred nanograms of RNA per tumor sample was heat
486 fragmented to a mean size of 200 base pairs in the presence of magnesium. Using random

487 primers, the RNA was used for first-strand cDNA synthesis, followed by second-strand synthesis
488 and A-tailing, adapter ligation, bead-based cleanup, and amplification of the library. After library
489 planning, the *IDT xGEN Exome Test Panel* was hybridized with samples. Streptavidin-coated
490 beads and target recovery were carried out, accompanied by amplification using the *KAPA HiFi*
491 library amplification package. The RNA libraries were sequenced on an *Illumina HiSeq 4000* using
492 patterned flow cell technology to achieve at least 50 million reads.
493
494 ***Detection of somatic variation on DNA sequencing data.*** The tumor and normal FASTQ files
495 were paired. For quality management measurement, FASTQ files were evaluated using FASTQC
496 and matched with Novoalign (Novocraft, Inc.)[83,84]. SAM files were generated and converted to
497 BAM files. The BAM files were sorted, and duplicates were marked. Single nucleotide variations
498 (SNVs) were called after alignment and sorting. For discovery of copy number alterations, the de-
499 duplicated BAM files and the VCF generated from the variant calling pipeline were processed to
500 computate read depth and variance of heterozygous germline SNVs between the tumor sample
501 and normal sample. Binary circular segmentation was introduced and segments with strongly
502 differential $\log_2$ ratios between the tumor and its comparator were chosen. From a combination of
503 differential coverage in segmented regions and estimation of stromal admixture provided by
504 analysis of heterozygous germline SNVs, an estimated integer copy number was determined
505
506 ***Microsatellite instability status.*** Probes for 43 microsatellite regions were developed using
507 *Tempus xT* assay[83]. Tumors were categorized into three groups by the MSI classification
508 algorithm as described by Tempus: microsatellite instability-high (MSI-H), microsatellite stable
509 (MSS) or microsatellite equivocal (MSE). MSI screening for paired tumor-normal patients used
510 reads mapped to the microsatellite loci with at least 5 bps flanking the microsatellite. The sample
511 was graded as MSI-H if there was a >70% chance of MSI-H classification. If the likelihood of MSI-
512 H status was 30-70%, the test findings were too ambiguous to interpret and those samples were

513 listed as MSE. If there was a <30% chance of MSI-H status, the sample was called MSS.
514 Additionally, IHC results were used to classify tumors into MSS or MSI molecular subtypes. Both
515 of these modalities were concordant and produced the same results.
516
517 ***Bulk RNA-seq and microarray analysis.*** We downloaded and pooled eight colorectal gene
518 expression datasets (GSE13067[85], GSE13294[85], GSE14333[86], GSE17536[87], GSE20916[88],
519 GSE33113[89], GSE35896[90], and GSE39582[14]), a pancreatic cancer dataset (GSE62165[91]) and
520 a non-small cell lung cancer dataset (GSE33532[92]) to validate our findings from the single cell
521 compartments by deconvoluting the bulk gene expression profiles into pseudo single-cell
522 resolutions. We used Affy (v1.64.0) for the data analysis and for exploration of Affymetrix
523 oligonucleotide array probe level data[93]. Batch correction was carried out using the
524 removeBatchEffect (v3.42.2) function of the LIMMA program and CMScaller for the CMS
525 classification (see below)[94]. Three datasets (GSE17536[87], GSE33113[89], and GSE39582[14]) were
526 utilized for clinical outcome analysis[94,95].
527
528 ***Correlation patterns in bulk gene expressions for CAF compartments.*** To identify the top
529 correlated CAF-marker genes within the combined eight CRC datasets, four bulk CRC gene
530 expression sets individually, pancreas cancer and lung cancer datasets. We first transformed the
531 bulk gene expression sets with $\log_2$ transformation. Next, marker genes with an average $\log_2$
532 FC>/= 0.5 and $p<0.05$ obtained from the single cell data of CAF-S1 and CAF-S4 compartments
533 were separately intersected with the bulk gene expression sets. Genes with an average
534 Spearman correlation score greater than 0.8 were kept as the CAF signatures within the bulk
535 gene expression. Heatmaps illustrating the correlation patterns within and between the CAF
536 compartments were prepared with the heatmap.2 function from ggplot package (v3.1.1) utilizing
537 the Pearson correlation coefficient. Heatmaps illustrating the correlation patterns within and

538 between the CAF compartments were prepared using the ggplot package (v3.1.1) utilizing the
539 Pearson correlation coefficient[96].
540
541 The Cox proportional hazard regression model was used to examine the significance of 20 cell
542 types from scRNA-seq in bulk expression data. Each cell type's marker genes with an average
543 logFC>1 and adjusted P<0.05 were intersected with the bulk expression datasets separately. We
544 only kept the marker genes with a high correlation with each other in bulk, which provides an
545 average correlation score of > 0.8. The average bulk expression of each cell type's remaining
546 marker genes was calculated and used in the hazard regression model as the representative of
547 this cell type. For analysis of relationships with patient outcome, univariate models were
548 calculated using Cox proportional hazard regression (coxph function from survival R package)[97].
549
550 ***Deconvoluting public bulk gene expression profiles into pseudo single-cell expressions.***
551 We used CIBERSORTx v1 to estimate composition of various cell populations in pooled eight
552 microarray datasets[52]. Signature gene matrices were created using the expression profiles of
553 49,859 cells as the reference single cell profile. We ran the 'hires' module with default parameters
554 except for the 'rmbatchBmode,' and the bulk-mode batch correction argument was set to true.
555 After the deconvolution process, we normalized the gene expressions according to the cell
556 fractions in each sample and calculated each gene's Z-transformed expression values. The
557 average normalized expression of each cell type across all samples was plotted with the
558 heatmap.3 R function of the GMD package (v0.3.3)[98]. A signature matrix highlighting marker
559 genes of the different cell types was prepared with a heatmap.2 R function of ggplot (v3.1.1). We
560 also applied the same parameters to deconvolute GSE14333[86], GSE17536[87], GSE33113[89], and
561 GSE39582[14] datasets individually.
562

563 ***Consensus molecular subtyping of colorectal cancer (CMS Classification).*** We used R
564 package CMScaller(v0.9.2), a nearest template prediction (NTP) algorithm, for the classification
565 of gene expression datasets[95]. We set the permutation number to 1000 to predict the CMS classes
566 of the samples in the GEO datasets with a p-value < 0.05. We ran CMScaller with default
567 parameters.
568
569 ***Continuous subtype discovery using scRNA-seq analysis.*** Bulk mRNA expression profiles of
570 the combined and batch adjusted eight GEO datasets (GSE13067[85], GSE13294[85], GSE14333[86],
571 GSE17536[87], GSE20916[88], GSE33113[89], GSE35896[90], and GSE39582[14]), composed of 1584
572 samples in total, were deconvoluted into the pseudo single-cell expression profiles via
573 CIBERSORTx utilizing the expression data consisting of 2o different cell types from our scRNA-
574 seq dataset[52]. We transformed the deconvoluted expression matrix with log2 transformation. The
575 principal components cluster subtype scores (PCSSs) of the CMS subtypes among the 1584
576 samples, were determined separately for each cell type using an algorithm published by Ma et
577 al[55]. To obtain the PCSSs, the average loading vectors were used. The results obtained for 20
578 cell types were projected on the first two PCSSs (PCSS1 and PCSS2) as they were validated by
579 Ma et al. in their analysis using 18 datasets. We also analyzed four datasets (GSE14333[86],
580 GSE17536[87], GSE33113[89], and GSE39582[14]) to independently confirm reproducibility of
581 continuous scores.
582
583 ***Statistics and reproducibility.*** All statistical analyses and graphs were created in R (v3.6.3) and
584 using a Python-based computational analysis tool. Schematic representations were made using
585 the Inkscape (https://inkscape.org/) software. Dim plots, bar plots and box plots were generated
586 using the dittoSeq (v1.1.7) package with default parameters[99]. Violin plots were generated using
587 the patchwork (v1.1.0) package and ggplot2 (v3.3.2) package in R with default parameters.
588 Heatmaps were generated using Morpheus.R with default parameters[100,101]. ANOVA and pair-

589 wised t-tests for the CMS classes across the deconvoluted expression profiles were performed in
590 R using the ggpubr R (v0.4.0) package[102]. The Box and Whisker plots were generated using the
591 boxplot function of the R base package at default parameters. The mean of the log_2 transformed
592 deconvoluted expression value of the samples in each CMS group was demonstrated with a
593 horizontal straight line within each box. The length of a boxplot corresponds to the interquartile
594 range (IQR), which is defined as the range between the first and third quartiles (Q1 and Q3),
595 whereas the whiskers are the upper and lower extreme values of the data (either data's extremum
596 values, or the Q3+1.5*IQR and Q1-1.5*IQR values, whichever was less extreme).
597
598 ***Survival analysis.*** Survival curves were obtained according to the Kaplan-Meier method survfit
599 (v3.2-7), and differences between survival distributions were assessed by Log-rank test. The
600 patients were divided into two groups (high/poor and low/good risk) according to their median
601 expression values (survminer (v0.4.8)). The surv_cutpoint function uses the maximally selected
602 rank statistics and implements standard methods for the approximation of the null distribution of
603 maximally selected rank statistics (maxstat (v0.7-25).
604
605 The proportional hazard assumption was tested to examine the fit of the model for survival of the
606 samples in four GEO datasets (GSE14333[86], GSE17536[87], GSE33113[89], and GSE39582[14]) with
607 respect to the deconvoluted bulk mRNA expressions. For analysis of the relationships with patient
608 outcome, multivariate models were calculated using the Cox proportional hazard regression
609 (coxph survival R package)[97].
610
611
612
613
614

615 **REFERENCES**
616

617 1. Cancer of the Colon and Rectum - Cancer Stat Facts. *SEER*
618 https://seer.cancer.gov/statfacts/html/colorect.html.
619 2. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut*
620 **66**, 683–691 (2017).
621 3. Osterman, E. *et al.* Recurrence risk after radical colorectal cancer surgery—less than
622 before, but how high is it? *Cancers* **12**, 3308 (2020).
623 4. Molinari, C. *et al.* Heterogeneity in colorectal cancer: a challenge for personalized medicine?
624 *Int J Mol Sci* **19**, (2018).
625 5. Xie, Y.-H., Chen, Y.-X. & Fang, J.-Y. Comprehensive review of targeted therapy for
626 colorectal cancer. *Signal Transduction and Targeted Therapy* **5**, 1–30 (2020).
627 6. Budinska, E. *et al.* Gene expression patterns unveil a new level of molecular heterogeneity
628 in colorectal cancer. *J Pathol* **231**, 63–76 (2013).
629 7. Deschoolmeester, V. *et al.* Tumor infiltrating lymphocytes: an intriguing player in the survival
630 of colorectal cancer patients. *BMC Immunol* **11**, 19 (2010).
631 8. Lee, W.-S., Park, S., Lee, W. Y., Yun, S. H. & Chun, H.-K. Clinical impact of tumor-
632 infiltrating lymphocytes for survival in stage II colon cancer. *Cancer* **116**, 5188–5199 (2010).
633 9. Perez, E. A. *et al.* Association of stromal tumor-infiltrating lymphocytes with recurrence-free
634 Survival in the N9831 adjuvant trial in patients with early-stage HER2-positive breast
635 cancer. *JAMA Oncology* **2**, 56–64 (2016).
636 10. Nearchou, I. P. *et al.* Spatial immune profiling of the colorectal tumor microenvironment
637 predicts good outcome in stage II patients. *npj Digital Medicine* **3**, 1–10 (2020).
638 11. Pagès, F. *et al.* International validation of the consensus Immunoscore for the classification
639 of colon cancer: a prognostic and accuracy study. *The Lancet* **391**, 2128–2139 (2018).
640 12. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human
641 colon and rectal cancer. *Nature* **487**, 330–337 (2012).
642 13. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular
643 phenotype and responses to therapy. *Nature Medicine* **19**, 619–625 (2013).
644 14. Marisa, L. *et al.* Gene Expression Classification of Colon Cancer into Molecular Subtypes:
645 Characterization, Validation, and Prognostic Value. *PLOS Medicine* **10**, e1001453 (2013).
646 15. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine*
647 **21**, 1350–1356 (2015).
648 16. Berg, I. van den *et al.* Improving clinical management of colon cancer through
649 CONNECTION, a nation-wide colon cancer registry and stratification effort (CONNECTION
650 II trial): rationale and protocol of a single arm intervention study. *BMC Cancer* **20**, 1–8
651 (2020).
652 17. Lenz, H.-J. *et al.* Impact of Consensus Molecular Subtype on Survival in Patients With
653 Metastatic Colorectal Cancer: Results From CALGB/SWOG 80405 (Alliance). *J Clin Oncol*
654 **37**, 1876–1885 (2019).
655 18. Stintzing, S. *et al.* Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and
656 first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306)
657 trial. *Ann Oncol* **30**, 1796–1803 (2019).
658 19. Mooi, J. K. *et al.* The prognostic impact of consensus molecular subtypes (CMS) and its
659 predictive effects for bevacizumab benefit in metastatic colorectal cancer: molecular
660 analysis of the AGITG MAX clinical trial. *Ann Oncol* **29**, 2240–2246 (2018).
661 20. Sveen, A., Cremolini, C. & Dienstmann, R. Predictive modeling in colorectal cancer: time to
662 move beyond consensus molecular subtypes. *Annals of Oncology* **30**, 1682–1685 (2019).
663 21. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters—
664 consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals*
665 *of Oncology* **30**, 520–527 (2019).

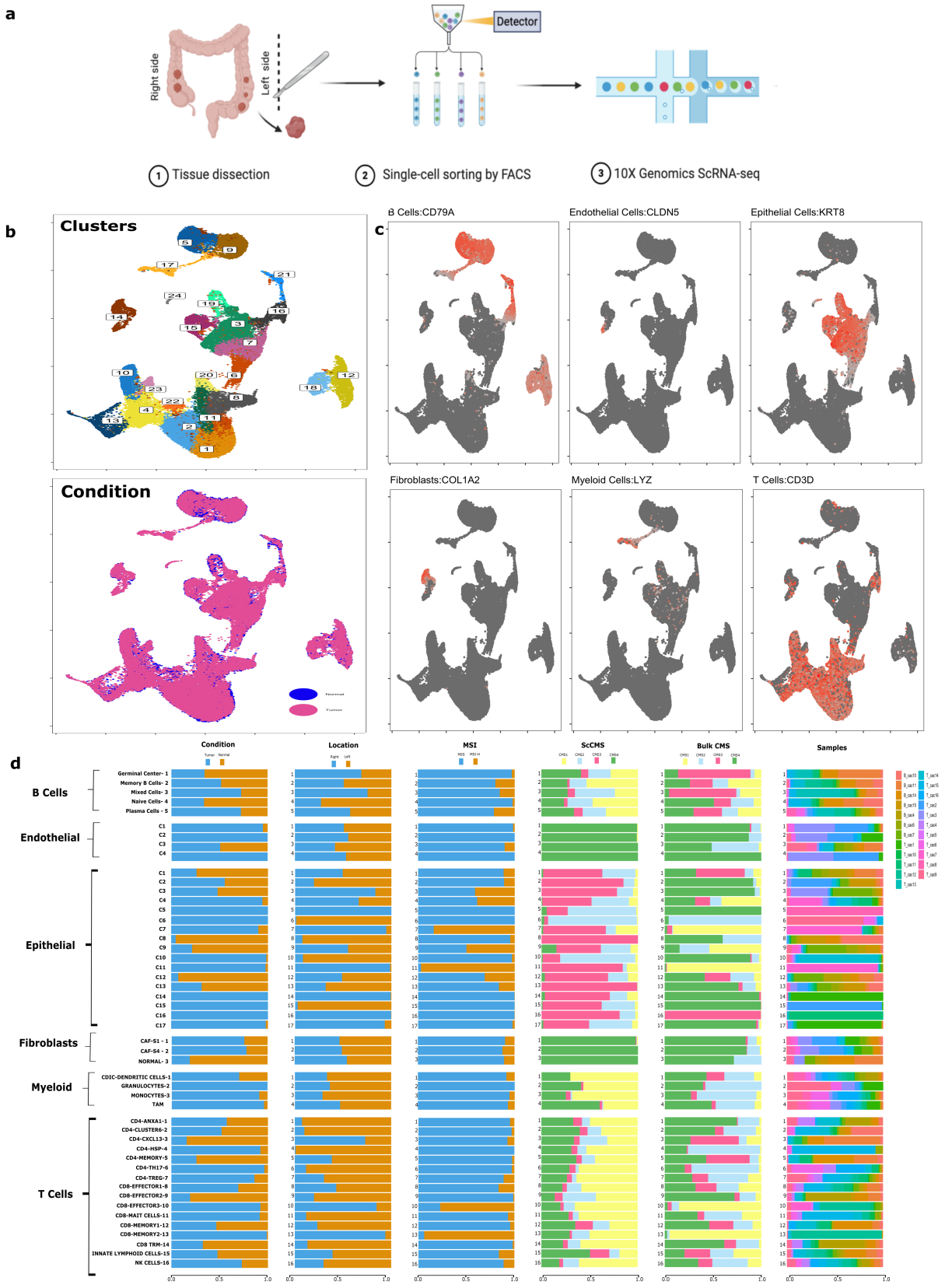
666 22. Khambata-Ford, S. *et al.* Expression of epiregulin and amphiregulin and K-ras mutation
667 status predict disease control in metastatic colorectal cancer patients treated with
668 cetuximab. *J Clin Oncol* **25**, 3230–3237 (2007).
669 23. Aderka, D., Stintzing, S. & Heinemann, V. Explaining the unexplainable: discrepancies in
670 results from the CALGB/SWOG 80405 and FIRE-3 studies. *The Lancet Oncology* **20**, e274–
671 e283 (2019).
672 24. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at
673 multiple resolutions. *Gigascience* **7**, (2018).
674 25. Zhang, L. *et al.* Single-cell analyses inform mechanisms of myeloid-targeted therapies in
675 colon cancer. *Cell* **181**, 442-459.e29 (2020).
676 26. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal
677 cancer. *Nature* **564**, 268–272 (2018).
678 27. Corridoni, D. *et al.* Single-cell atlas of colonic CD8 + T cells in ulcerative colitis. *Nature*
679 *Medicine* **26**, 1480–1490 (2020).
680 28. Oh, D. Y. *et al.* Intratumoral CD4+ T Cells Mediate Anti-tumor Cytotoxicity in Human
681 Bladder Cancer. *Cell* **181**, 1612-1625.e13 (2020).
682 29. Lee, H.-O. *et al.* Lineage-dependent gene expression programs influence the immune
683 landscape of colorectal cancer. *Nature Genetics* **52**, 594–603 (2020).
684 30. Mills, J. C. & Sansom, O. J. Reserve stem cells: Differentiated cells reprogram to fuel repair,
685 metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci Signal* **8**, re8 (2015).
686 31. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
687 microenvironment. *Nat Med* **24**, 1277–1289 (2018).
688 32. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nature Medicine*
689 **26**, 1271–1279 (2020).
690 33. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
691 collection. *Cell Syst* **1**, 417–425 (2015).
692 34. Jamal-Hanjani, M. *et al.* Tracking the evolution of non-small-cell lung cancer. *N Engl J Med*
693 **376**, 2109–2121 (2017).
694 35. Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of
695 subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* **10**,
696 2750 (2019).
697 36. Lim, S. B. *et al.* Addressing cellular heterogeneity in tumor and circulation for refined
698 prognostication. *Proc Natl Acad Sci U S A* **116**, 17957–17962 (2019).
699 37. Wang, R. *et al.* Single-cell dissection of intratumoral heterogeneity and lineage diversity in
700 metastatic gastric adenocarcinoma. *Nat Med* **27**, 141–151 (2021).
701 38. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray
702 and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
703 39. Sathe, A. *et al.* Single-cell genomic characterization reveals the cellular reprogramming of
704 the gastric tumor microenvironment. *Clin Cancer Res* **26**, 2640–2653 (2020).
705 40. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
706 pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
707 41. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat*
708 *Methods* **14**, 979–982 (2017).
709 42. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon
710 tumors. *Nat Biotechnol* **29**, 1120–1127 (2011).
711 43. Vermeulen, L. *et al.* Single-cell cloning of colon cancer stem cells reveals a multi-lineage
712 differentiation capacity. *PNAS* **105**, 13427–13432 (2008).
713 44. Monk, M. & Holding, C. Human embryonic genes re-expressed in cancer cells. *Oncogene*
714 **20**, 8085–8091 (2001).
715 45. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–
716 767 (1990).

717 46. Costa, A. *et al.* Fibroblast heterogeneity and immunosuppressive environment in human
718 breast cancer. *Cancer Cell* **33**, 463-479.e10 (2018).
719 47. Bartoschek, M. *et al.* Spatially and functionally distinct subclasses of breast cancer-
720 associated fibroblasts revealed by single cell RNA sequencing. *Nat Commun* **9**, 5150
721 (2018).
722 48. Kieffer, Y. *et al.* Single-cell analysis reveals fibroblast clusters linked to immunotherapy
723 resistance in cancer. *Cancer Discov* **10**, 1330–1351 (2020).
724 49. Wu, S. Z. *et al.* Stromal cell diversity associated with immune evasion in human triple-
725 negative breast cancer. *The EMBO Journal* **39**, (2020).
726 50. Sadanandam, A. *et al.* Reconciliation of classification systems defining molecular subtypes
727 of colorectal cancer. *Cell Cycle* **13**, 353–357 (2014).
728 51. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct
729 subtype and develops from serrated precursor lesions. *Nat Med* **19**, 614–618 (2013).
730 52. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues
731 with digital cytometry. *Nature Biotechnology* **37**, 773–782 (2019).
732 53. Dunne, P. D. *et al.* Challenging the cancer molecular stratification dogma: intratumoral
733 heterogeneity undermines consensus molecular subtypes and potential piagnostic value in
734 colorectal cancer. *Clin Cancer Res* **22**, 4095–4104 (2016).
735 54. Roepman, P. *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit,
736 deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of*
737 *Cancer* **134**, 552–562 (2014).
738 55. Ma, S. *et al.* Continuity of transcriptomes among colorectal cancer subtypes based on meta-
739 analysis. *Genome Biology* **19**, 142 (2018).
740 56. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms.
741 https://link.springer.com/article/10.1007/s10852-005-9022-1.
742 57. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with
743 confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
744 58. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC*
745 *Bioinformatics* **11**, 367 (2010).
746 59. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular
747 heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708–718 (2017).
748 60. Liu, T. *et al.* Cancer-associated fibroblasts: an emerging target of anti-cancer
749 immunotherapy. *J Hematol Oncol* **12**, 86 (2019).
750 61. Barnett, R. M. & Vilar, E. Targeted therapy for cancer-associated fibroblasts: are we there
751 yet? *JNCI: Journal of the National Cancer Institute* **110**, 11–13 (2018).
752 62. Gascard, P. & Tlsty, T. D. Carcinoma-associated fibroblasts: orchestrating the composition
753 of malignancy. *Genes Dev.* **30**, 1002–1019 (2016).
754 63. Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated
755 fibroblasts. *Nat Rev Cancer* **20**, 174–186 (2020).
756 64. Roth, A. D. *et al.* Prognostic role of KRAS and BRAF in stage II and III resected colon
757 cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00
758 trial. *J Clin Oncol* **28**, 466–474 (2010).
759 65. Laurent-Puig, P. *et al.* Colon cancer molecular subtype intratumoral heterogeneity and its
760 prognostic impact: An extensive molecular analysis of the PETACC-8. *Annals of Oncology*
761 **29**, viii18 (2018).
762 66. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902.e21
763 (2019).
764 67. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet detection in single-
765 cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* **8**, 329-337.e4 (2019).
766 68. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome*
767 *Biol* **11**, R106 (2010).

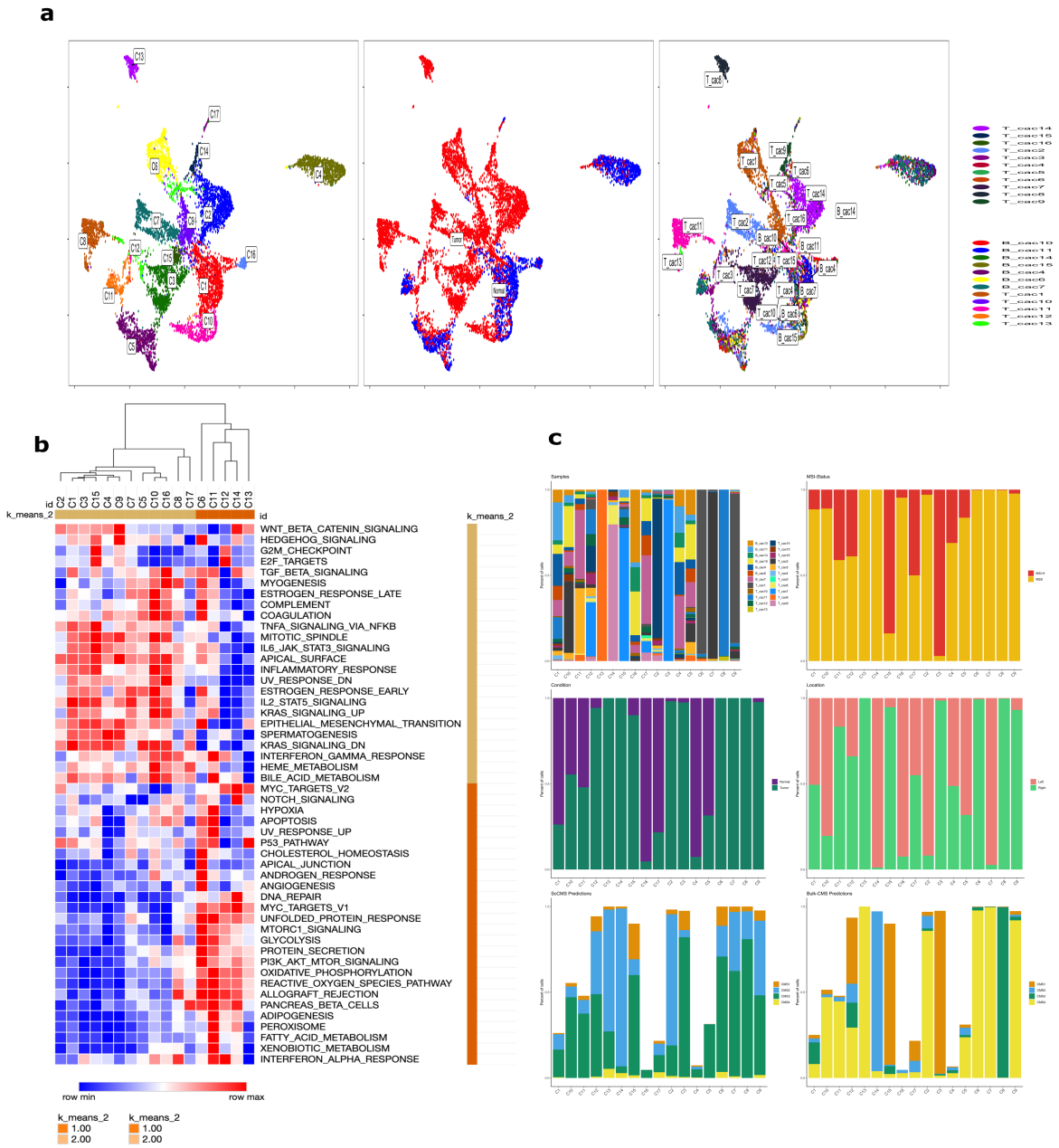
69. Thompson, B. Canonical correlation analysis. in *Encyclopedia of Statistics in Behavioral Science* (American Cancer Society, 2005). doi:10.1002/0470013192.bsa068.
70. Jolliffe, I. Principal component analysis. in *International Encyclopedia of Statistical Science* (ed. Lovric, M.) 1094–1096 (Springer, 2011). doi:10.1007/978-3-642-04898-2_455.
71. Maaten, L. van der. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014).
72. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018).
73. Zilionis, R. *et al.* Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity* **50**, 1317-1334.e10 (2019).
74. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).
75. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med* **24**, 978–985 (2018).
76. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* **10**, 4706 (2019).
77. Nirschl, C. J. *et al.* IFNγ-dependent tissue-Immune homeostasis Is co-opted in the tumor microenvironment. *Cell* **170**, 127-141.e15 (2017).
78. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285 (2020).
79. Shi, Z. *et al.* More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discov* **5**, 64 (2019).
80. Ramesh, A. *et al.* A pathogenic and clonally expanded B cell transcriptome in active multiple sclerosis. *Proc Natl Acad Sci U S A* **117**, 22932–22943 (2020).
81. Puram, S. V. *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611-1624.e24 (2017).
82. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309–315 (2017).
83. Beaubier, N. *et al.* Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. *Oncotarget* **10**, 2384–2396 (2019).
84. Andrews, S. *FastQC: a quality control tool for high throughput sequence data.* (Babraham Institute, 2012).
85. Jorissen, R. N. *et al.* DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* **14**, 8061–8069 (2008).
86. Jorissen, R. N. *et al.* Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin Cancer Res* **15**, 7642–7651 (2009).
87. Smith, J. J. *et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* **138**, 958–968 (2010).
88. Skrzypczak, M. *et al.* Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One* **5**, (2010).
89. Kemper, K. *et al.* Mutations in the Ras-Raf Axis underlie the prognostic value of CD133 in colorectal cancer. *Clin Cancer Res* **18**, 3132–3141 (2012).
90. Schlicker, A. *et al.* Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics* **5**, 66 (2012).
91. Janky, R. *et al.* Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma. *BMC Cancer* **16**, 632 (2016).

92. Meister, M. *et al.* Intra-tumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer. *Journal of Bioinformatics Research Studies* **1**, (2014).
93. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
94. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
95. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep* **7**, 16618 (2017).
96. Wickham, H. *ggplot2*. (Springer).
97. Therneau, T. *A package for survival analysis in R*. (2020).
98. Zhao, X., Valen, E., Parker, B. J. & Sandelin, A. Systematic clustering of transcription start site landscapes. *PLOS ONE* **6**, e23409 (2011).
99. Bunis, D. G., Andrews, J., Fragiadakis, G. K., Burt, T. D. & Sirota, M. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1011.
100. *morpheus: Interactive heat maps using 'morpheus.js' and 'htmlwidgets'*. (2021).
101. Pedersen, T. L. *Patchwork: The Composer of Plots. R*. (2020).
102. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. (Based Publication, 2020).

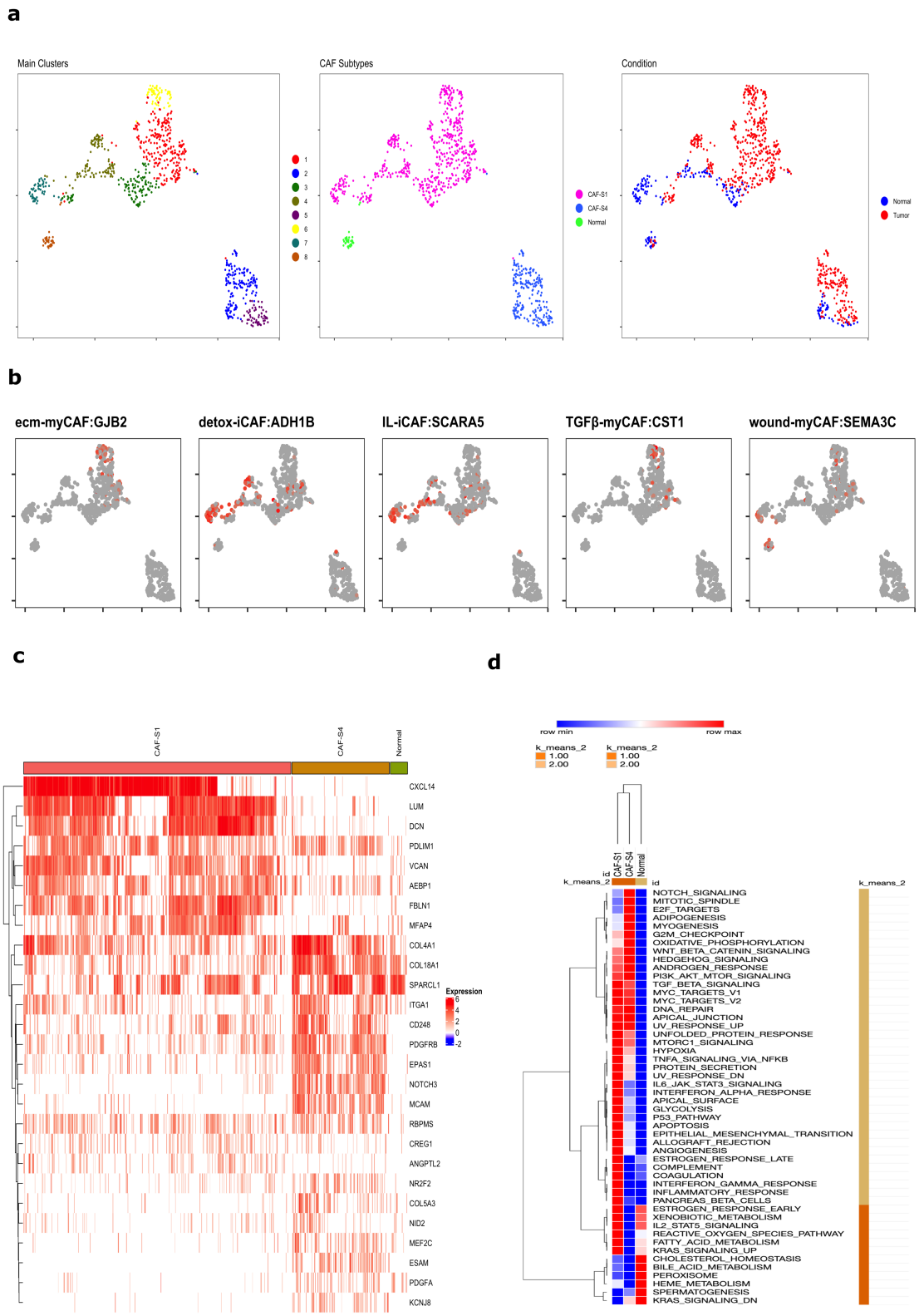
FIGURES


a
Right side
Left side
Detector
1 Tissue dissection
2 Single-cell sorting by FACS
3 10X Genomics ScRNA-seq
b
Clusters
Condition
c
B Cells:CD79A
Endothelial Cells:CLDN5
Epithelial Cells:KRT8
Fibroblasts:COL1A2
Myeloid Cells:LYZ
T Cells:CD3D
d
Condition
Location
MSI
ScCMS
Bulk CMS
Samples
B Cells
Endothelial
Epithelial
Fibroblasts
Myeloid
T Cells

856 **Figure 1. Identification and clustering of single cells**. **a,** Workflow of sample collection,
857 sorting, and sequencing (methods contain full description for each step). **b,** UMAP
858 characterization of the 49,859 cells profiled. Coloring demonstrates clusters, tumor vs. non-
859 malignant sample origin (condition), and individual sample origin. **c,** Identification of various cell
860 types based on expression of specified marker genes. **d,** Characterization of the proportion of
861 cell types identified in tumor vs. non-malignant tissue, sidedness (right vs. left), microsatellite
862 instability (MSI) status, single-cell Consensus Molecular Subtypes (scCMS) classification,
863 Consensus Molecular Subtypes (CMS) of bulk RNA-seq data, and origin of sample. The
864 transcription counts of tumor and normal tissue cell types are demonstrated at the bottom with
865 boxplot representation. The graph represents total clusters and cell types identified after re-
866 clustering of each cell compartment depicting global heterogeneous landscape of colorectal
867 cancers.

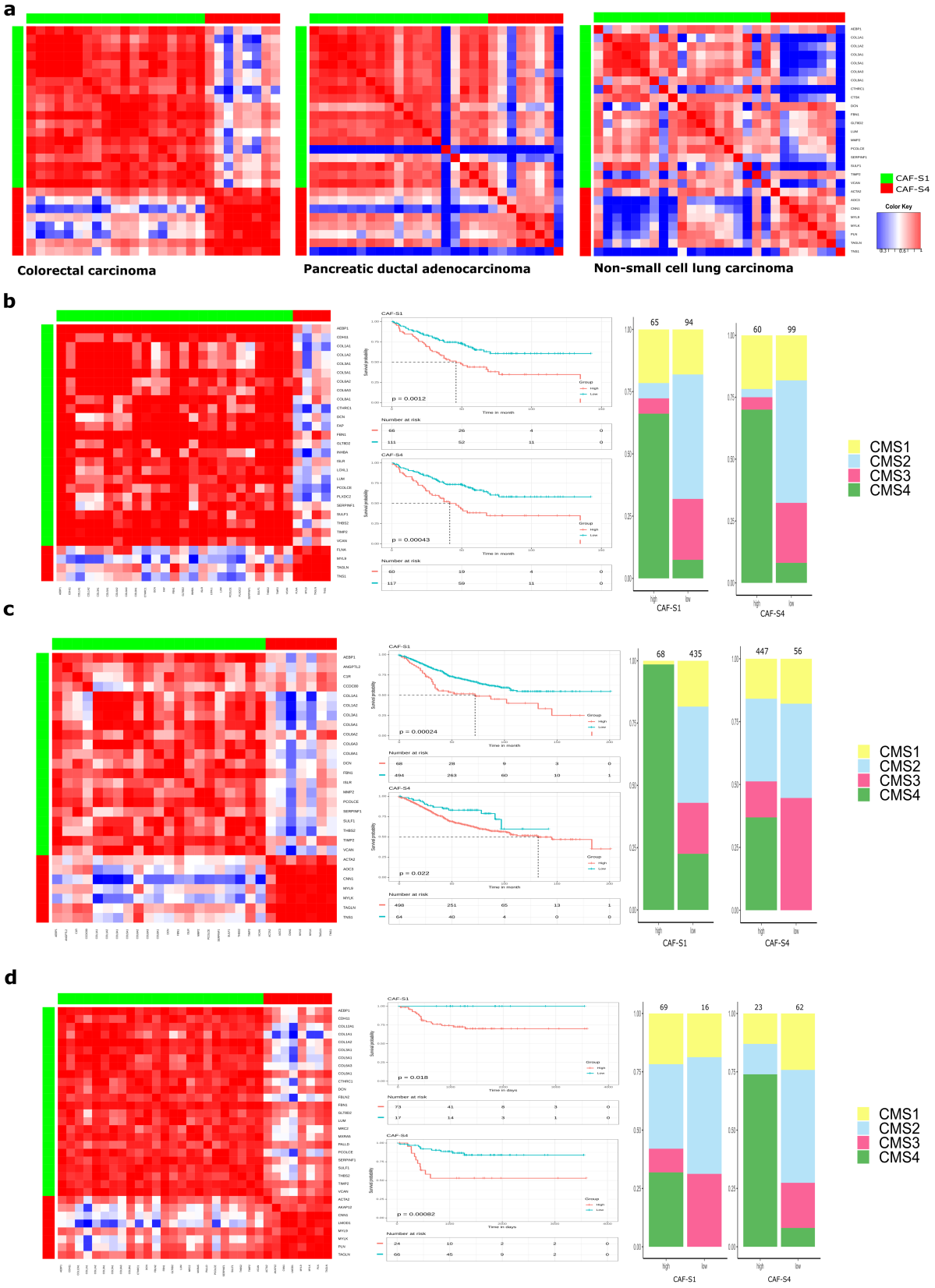

a
b
c
d
id
k_means_2
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
G2M_CHECKPOINT
E2F_TARGETS
TGF_BETA_SIGNALING
MYOGENESIS
ESTROGEN_RESPONSE_LATE
COMPLEMENT
COAGULATION
TNFA_SIGNALING_VIA_NFKB
MITOTIC_SPINDLE
IL6_JAK_STAT3_SIGNALING
APICAL_SURFACE
INFLAMMATORY_RESPONSE
UV_RESPONSE_DN
ESTROGEN_RESPONSE_EARLY
IL2_STAT5_SIGNALING
KRAS_SIGNALING_UP
EPITHELIAL_MESENCHYMAL_TRANSITION
SPERMATOGENESIS
KRAS_SIGNALING_DN
INTERFERON_GAMMA_RESPONSE
HEME_METABOLISM
BILE_ACID_METABOLISM
MYC_TARGETS_V2
NOTCH_SIGNALING
HYPOXIA
APOPTOSIS
UV_RESPONSE_UP
P53_PATHWAY
CHOLESTEROL_HOMEOSTASIS
APICAL_JUNCTION
ANDROGEN_RESPONSE
ANGIOGENESIS
DNA_REPAIR
MYC_TARGETS_V1
UNFOLDED_PROTEIN_RESPONSE
MTORC1_SIGNALING
GLYCOLYSIS
PROTEIN_SECRETION
PI3K_AKT_MTOR_SIGNALING
OXIDATIVE_PHOSPHORYLATION
REACTIVE_OXYGEN_SPECIES_PATHWAY
ALLOGRAFT_REJECTION
PANCREAS_BETA_CELLS
ADIPOGENESIS
PEROXISOME
FATTY_ACID_METABOLISM
XENOBIOTIC_METABOLISM
INTERFERON_ALPHA_RESPONSE
row min
row max
k_means_2
1.00
2.00
Samples
MSI-Status
Condition
Location
ScCMS Predictions
Bulk-CMS Predictions
states
Location
MSI
Sc-CMS
Bulk-CMS
Component 1
Component 2

869 **Figure 2. Reclustering and characterization of the epithelial compartment. a,** UMAP of
870 tumor and non-malignant epithelial reclustering demonstrating 17 distinct clusters. **b,** Heatmap
871 of Hallmark pathway analysis within the epithelial cell compartment. **c,** Bar chart representation
872 of cell proportions by sample, tissue type, MSI status, colonic location of sample, scCMS score,
873 and bulk CMS score. **d,** Trajectory analysis of cells colored by colonic location, scCMS, MSI
874 status, and bulk CMS status.


a
Main Clusters
CAF Subtypes
Condition
1
2
3
4
5
6
7
8
CAF-S1
CAF-S4
Normal
Normal
Tumor
b
ecm-myCAF:GJB2
detox-iCAF:ADH1B
IL-iCAF:SCARA5
TGFβ-myCAF:CST1
wound-myCAF:SEMA3C
c
CAF-S1
CAF-S4
Normal
CXCL14
LUM
DCN
PDLIM1
VCAN
AEBP1
FBLN1
MFAP4
COL4A1
COL18A1
SPARCL1
ITGA1
CD248
PDGFRB
EPAS1
NOTCH3
MCAM
RBPMS
CREG1
ANGPTL2
NR2F2
COL5A3
NID2
MEF2C
ESAM
PDGFA
KCNJ8
Expression
6
4
2
0
-2
d
row min
row max
k_means_2
1.00
2.00
id
CAF-S1
CAF-S4
Normal
NOTCH_SIGNALING
MITOTIC_SPINDLE
E2F_TARGETS
ADIPOGENESIS
MYOGENESIS
G2M_CHECKPOINT
OXIDATIVE_PHOSPHORYLATION
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
ANDROGEN_RESPONSE
PI3K_AKT_MTOR_SIGNALING
TGF_BETA_SIGNALING
MYC_TARGETS_V1
MYC_TARGETS_V2
DNA_REPAIR
APICAL_JUNCTION
UV_RESPONSE_UP
UNFOLDED_PROTEIN_RESPONSE
MTORC1_SIGNALING
HYPOXIA
TNFA_SIGNALING_VIA_NFKB
PROTEIN_SECRETION
UV_RESPONSE_DN
IL6_JAK_STAT3_SIGNALING
INTERFERON_ALPHA_RESPONSE
APICAL_SURFACE
GLYCOLYSIS
P53_PATHWAY
APOPTOSIS
EPITHELIAL_MESENCHYMAL_TRANSITION
ALLOGRAFT_REJECTION
ANGIOGENESIS
ESTROGEN_RESPONSE_LATE
COMPLEMENT
COAGULATION
INTERFERON_GAMMA_RESPONSE
INFLAMMATORY_RESPONSE
PANCREAS_BETA_CELLS
ESTROGEN_RESPONSE_EARLY
XENOBIOTIC_METABOLISM
IL2_STAT5_SIGNALING
REACTIVE_OXYGEN_SPECIES_PATHWAY
FATTY_ACID_METABOLISM
KRAS_SIGNALING_UP
CHOLESTEROL_HOMEOSTASIS
BILE_ACID_METABOLISM
PEROXISOME
HEME_METABOLISM
SPERMATOGENESIS
KRAS_SIGNALING_DN

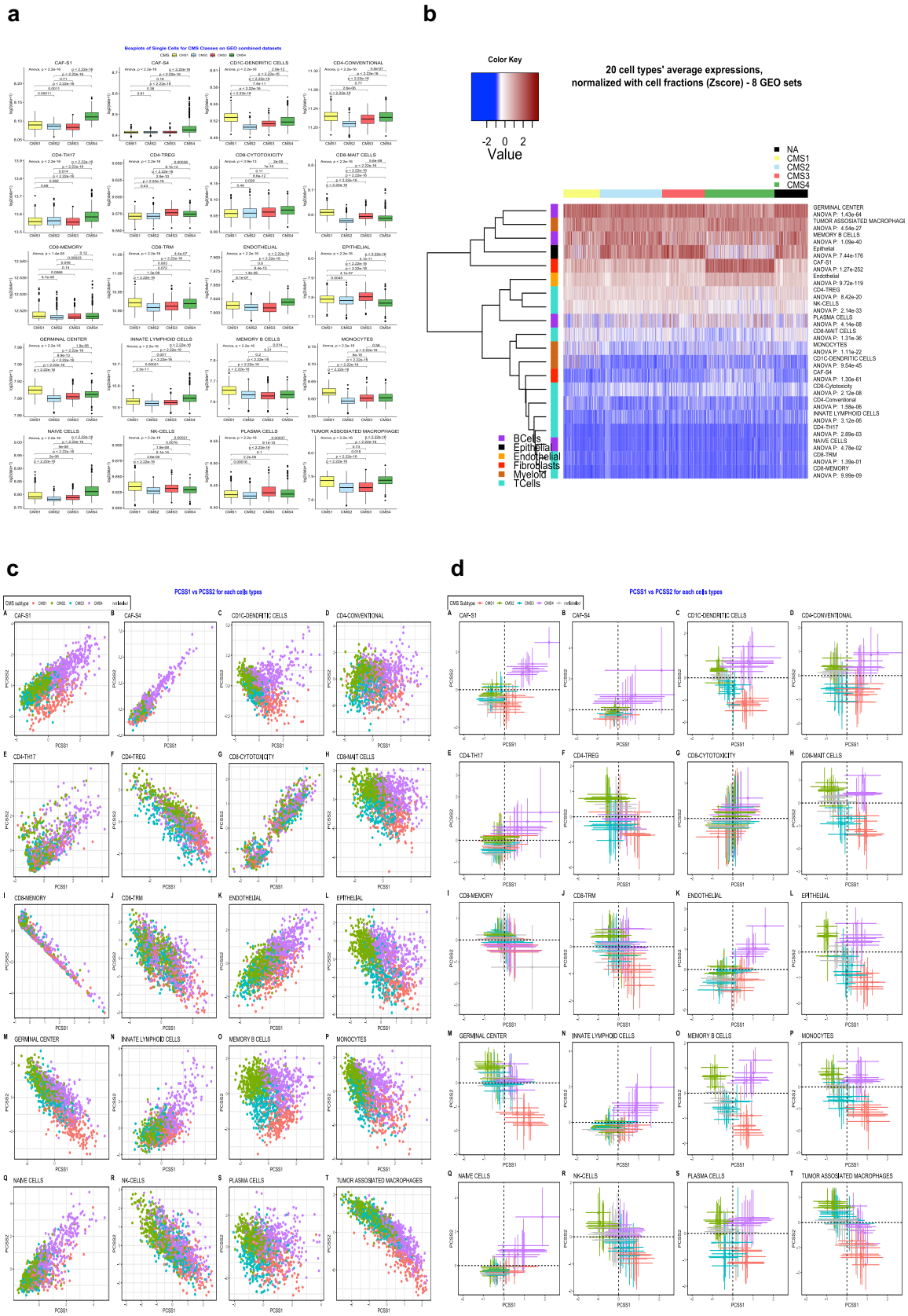
875

876 **Figure 3. Fibroblast clusters in colon and colorectal tumors. a,** UMAP of 819 fibroblasts
877 colored by distinct clusters, CAF status, tissue status and origin of sample. UMAP fibroblasts
878 colored by specific CAF-S1 subtypes. **b,** UMAP color-coded for marker genes for five CAF-S1
879 subtypes as indicated. **c,** Heatmap showing the variable expression of fibroblast specific marker
880 genes across CAF-S1, CAF-S4, and normal fibroblasts. **d,** Heatmap of Hallmark pathway
881 analysis of CAF-S1, CAF-S4, and normal cluster.


a
Colorectal carcinoma
Pancreatic ductal adenocarcinoma
Non-small cell lung carcinoma
CAF-S1
CAF-S4
Color Key
b
CAF-S1
p = 0.0012
Number at risk
CAF-S4
p = 0.00043
Number at risk
65
94
60
99
CAF-S1
CAF-S4
CMS1
CMS2
CMS3
CMS4
c
CAF-S1
p = 0.00024
Number at risk
CAF-S4
p = 0.022
Number at risk
68
435
447
56
CAF-S1
CAF-S4
CMS1
CMS2
CMS3
CMS4
d
CAF-S1
p = 0.018
Number at risk
CAF-S4
p = 0.00082
Number at risk
69
16
23
62
CAF-S1
CAF-S4
CMS1
CMS2
CMS3
CMS4

882

Figure 4. Correlation of CAF-S1 and CAF-S4 gene profiles across human bulk transcriptomic data. a, Pearson's correlation of genes from CAF-S1 and CAF-S4 profiles in colorectal cancer (n= 1584; CAF-S1 and CAF-S4 r > 0.8), pancreatic cancer (n= 118; CAF-S1 r = 0.70, CAF-S4 r= 0.60), non-small cell lung cancer (n = 80; CAF-S1 r = 0.69, CAF-S4 r = 0.67). **b-d,** Pearson correlation plots, Kaplan-Meyer survival curves, and bar plots of CMS status assessing CAF expression in individual CRC datasets. Plots b-c are generated from single GEO datasets; GSE17536 (n = 177), GSE39582 (n= 585) and GSE33113 (n= 96), respectively. Note: High CAF-S1 and CAF-S4 gene signatures are associated with poor survival across all CMS subtypes. r= coefficient correlation.


a
Boxplots of Single Cells for CMS Classes on GEO combined datasets
b
Color Key
-2 0 2
Value
20 cell types' average expressions,
normalized with cell fractions (Zscore) - 8 GEO sets
NA
CMS1
CMS2
CMS3
CMS4
BCells
Epithelial
Endothelial
Fibroblasts
Myeloid
TCells
c
PCSS1 vs PCSS2 for each cells types
d
PCSS1 vs PCSS2 for each cells types

914

915 **Figure 5. Average cell type abundance from eight pooled CRC datasets and sorted by**
916 **bulk CMS status. a,** Boxplots show the distribution of cell types within tumors with varying CMS
917 status. The whiskers depict the 1.5 x IQR. The p-values for one-way ANOVA are shown in the
918 figure. **b,** Deconvolution heatmap of different cell types by average expression using
919 CIBERSORTx demonstrating cell type distribution (based on individual datasets) within each
920 CMS category. **c,** All 20 cell types show no to little separation reported by CMS. **d,** All cell types
921 projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Markers
922 are colored by the bulk CMS status. Note that the cell types largely form a continuum along
923 CMS status and are not clustered in discrete quadrants separate from one another. Cells are
924 colored by bulk CMS status accordingly to origin of sample.