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 that significantly stratifies overall survival in more than 1,500 CRC patients with bulk transcriptomic data. 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.

68 **Main**

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

80

81 Bulk gene expression analyses by several independent groups have identified distinct CRC
82 subtypes[13–15]. Reflecting both the tumor and TME, an international consortium published the
83 Consensus Molecular Subtypes (CMS), which proposed four distinct subtypes of CRC[13,14,16].
84 Unfortunately, the association between CMS and meaningful therapeutic response to specific
85 agents have shown inconsistent results across studies and CMS lacks a concordance between
86 primary and metastatic CRC tumors, limiting its utility thus for in clinical decision making[17–24]. As
87 a result, an improved CMS classification or an alternative classification system is required to
88 improve clinical utility.

89

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

94 (CAF) population. The CAF population's clinical and prognostic significance became apparent
95 when CAF signatures were applied to large, independent CRC transcriptomic cohorts.
96
97 **RESULTS**
98 We profiled sixteen primary colon tissue samples and eight adjacent non-malignant tissues (24 in
99 total) using droplet-based, scRNA-seq. Altogether, we captured and retained 49,589 single cells
100 after performing quality control for downstream analysis (**Fig. 1a, supplementary table 1**). All
101 scRNA-seq data were merged and normalized to identify robust discrete clusters of epithelial cells
102 (*EPCAM+, KRT8+,* and *KRT18+*), fibroblasts (*COL1A2+*), endothelial cells (*CD31+*), T cells
103 (*CD3D+),* B cells *(CD79A+),* and myeloid cells (*LYZ+*) using canonical marker genes. Additionally,
104 each cell type compartment was analyzed separately. Cluster v0.4.1 and manual review of
105 differentially expressed genes in each subcluster were studied to choose the best cluster
106 resolution without cluster destabilization (see methods)[25]. Cell population designation was
107 chosen by specific gene expression, and SingleR was also utilized for unbiased cell type
108 recognition (see methods)[26–29].
109
110 In addition to cancer cells, we identified diverse TME cell phenotypes, including fibroblasts
111 subsets (*CAF-S1* and *CAF-S4*), endothelial cells, CD4+ subsets (*naïve/memory, Th17,* and
112 *Tregs*), CD8+ subsets (*naïve/memory, cytotoxic, tissue-resident memory,* and *Mucosa-*
113 *Associated Invariant (MAIT) cells*), NK cells, innate lymphoid cell (ILC) types, B cell phenotypes
114 (*naïve, memory, germinal center,* and *plasma cells*), and monocyte lineage phenotypes (*C1DC+*
115 *dendritic cells, proinflammatory monocyte*s [*IL1B, IL6, S100A8,* and *S100A9]*), and M2 polarized
116 anti-inflammatory [*CD163, SEPP1, APOE,* and *MAF]*), tumor-associated macrophages (TAMs)
117 (**Fig. 1b-d, Extended Data Figs. 1-3, and Extended Data Tables 1-4**)[26–29].
118

119 For validation, we additionally profiled 31,383 high-quality, single cells from an independent cohort
120 using stringent criteria to corroborate our findings (see methods)[30]. Thus, a total of 81,242 high-
121 quality cells were profiled to produce a single-cell map of 40 colorectal cancer patients. The results
122 of the primary CRC cohort (49,859 single-cells) are available at the Colon Cancer Atlas
123 (www.websiteinprogress.com).
124
125 **Malignant colon cancer reveals tumor epithelial cell subclonal heterogeneity and**
126 **stochastic behavior.**
127
128 We detected 8,965 tumor and benign epithelial cells (*EPCAM+, KRT8+,* and *KRT18+*) and, on
129 reclustering, produced 17 epithelial clusters (designated C1 to C17) (**Fig. 2a and Supplementary**
130 **Table 2**). Clusters were chiefly influenced by colonic epithelial markers, including those for
131 stemness (*LGR5, ASCL2, OLFM4,* and *STMN1*), enterocytes (*FABP1* and *CA2*), goblet cells
132 (*ZG16, MUC2, SPINK4,* and *TFF3*), and enteroendocrine cells (*PYY* and *CHGA*)
133 (**Supplementary Fig. 2a and Supplementary Table 2**). Some tumor-derived clusters (C7, C8,
134 and C13) did not express known colon epithelial markers, potentially representing a de-
135 differentiated state of plasticity (**Supplementary Fig. 2b**)[31]. Each distinct tumor-derived cluster
136 was predominantly patient-specific, reflecting a high degree of inter-patient tumoral cell
137 heterogeneity. In contrast, epithelial populations derived from non-malignant tissue samples from
138 multiple patients clustered together, a pattern observed in previous studies confirming both
139 normal tissue homeostasis and limited sample batch effects (**Fig. 2a**)[32,33].
140
141 We next aimed to identify gene expression programs shared across these clusters using hallmark
142 pathway analysis[34]. A strong overlap was observed for multiple pathways such as activation of
143 inflammatory, epithelial-mesenchymal transformation (EMT), immune response, and metabolic
144 pathways (**Fig. 2b**). Interestingly, high microsatellite instability (MSI-H) and microsatellite stable

145 (MSS) CRC tumors, considered clinically separate entities, demonstrated similar pathway
146 program activation within the tumor epithelial populations. Each cluster also showed activation of
147 unique pathways such as activation of angiogenesis (C6), Notch signaling (C14), apoptosis (C11),
148 hypoxia (C11), and TNF signaling dysregulation (C15), among others. However, MSI-H tumors
149 differed from MSS tumors based on immune cell infiltration (**Extended Data Fig. 1**).
150
151 Since intratumoral heterogeneity is recognized as a key mechanism contributing to drug
152 resistance, cancer progression, and recurrence, we next focused on dissecting potential
153 transcriptomic states to identify heterogeneity within each patient's tumor[35–38]. We found that each
154 tumor specimen contained 2-10 distinct tumor epithelial clusters (**Fig. 2c**). Gene set variation
155 analysis (GSVA) was performed on cells from individual tumor samples and illustrated the sub-
156 clonal transcriptomic heterogeneity within each specimen (**Supplementary Fig. 2c**)[39]. Clusters
157 identified in individual pathway analysis demonstrated the up- or down-regulation of crucial
158 metabolic and oncogenic pathways between samples, suggesting wide phenotype variations
159 between cells from the same tumor[40].
160
161 Given the evidence of intratumoral epithelial heterogeneity, we next performed trajectory
162 inference using pseudotime analysis to identify potential alignments or lineage relationships (i.e.,
163 right versus left-sided CRC), CMS classification, or MSI status
164 (**Fig. 2d**)[41,42]. This analysis also served as a control for inter-patient genomic heterogeneity and
165 provided an orthogonal strategy to confirm the transcriptomic trends we identified. We detected
166 five molecular states (S1n/t to S5n/t) with malignant and normal epithelial cells intermixed along
167 a joint transcriptional trajectory. This observation is consistent with prior studies demonstrating
168 that CRC cells recapitulate normal colon epithelia's multilineage differentiation process as each
169 transcriptional state's pathway activation in both normal and tumor cells was related to normal

170 colon epithelial function of nutrient absorption or maintaining colon homeostasis (**Supplementary**
171 **Fig. 3 and Supplementary Table 3**)[43,44].
172
173 Additionally, tumor cells showed upregulation of embryogenesis (S2t), consistent with previous
174 findings that tumor cells revert to their embryological states in cancer development
175 (**Supplementary Table 4**)[45]. Interestingly, there were no significant associations with anatomic
176 location, CMS classification, or MSI status within our dataset or an independent dataset of 31,383
177 single cells (**Supplementary Fig. 4**)[30]. Hence, in our analysis, CRC development mainly
178 represents a hijacking of the normal epithelial differentiation program, coupled with the acquisition
179 of additional cancer-related and embryogenic pathways (**Supplementary Fig. 3**)[46].
180
181 **CRC-associated fibroblasts in the tumor microenvironment exhibit diverse phenotypes,**
182 **and specific subtypes are associated with poor prognosis.**
183
184 We next focused on CRC TME subpopulations. High-quality 819 fibroblasts were re-clustered into
185 eight clusters, and then phenotypically classified into two major subtypes to assess for further
186 CAF heterogeneity. These phenotypic subtypes were found to be immunomodulatory CAF-S1
187 (*PDGFRA*+ and *PDPN*+) and prometastatic CAF-S4 (*RGS5*+ and *MCAM*+) (**Fig. 3 and**
188 **Supplementary Table 4**)[47]. This fibroblast cluster dichotomy was also observed in the
189 independent CRC patient scRNA-seq dataset of 31,383 cells (**Supplementary Fig. 5**)[30].
190
191 The CAF-S1 and CAF-S4 subtypes showed striking resemblances to the mCAF (extracellular
192 matrix) and vCAF (vascular) fibroblast subtypes, respectively, as previously described in a mouse
193 breast cancer model[48]. Most clusters were found in multiple patients, albeit in varying proportions,
194 signifying shared patterns in CAF transcriptomic programs between patients. Fibroblasts derived
195 from MSI-H tumors were distributed similarly throughout these clusters.

196

197 CAF-S1 exhibited high chemokine expressions such as *CXCL1, CXCL2, CXCL12, CXCL14,* and
198 immunomodulatory molecules including *TNFRSF12A* (**Supplementary Fig. 6**). Additionally,
199 CAF-S1 expressed extracellular matrix genes including matrix-modifying enzymes (*LOXL1* and
200 *LOX*)[48]. To determine this population's functional significance, we compared the CAF-S1
201 population transcriptomes to those described recently in breast cancer, lung cancer, and head
202 and neck cancer[49]. We recovered five CAF subtypes, within the CAF-S1 population, including
203 *ecm-myCAF* (extracellular; *GJB*+), *IL-iCAF* (growth factor, *TN*F and interleukin pathway;
204 *SCARA5*+), *detox-iCAF* (detoxification and inflammation; *ADH1B*+), *wound-myCAF* (collagen
205 fibrils and wound healing; *SEMA3C*), and *TGFβ-myCAF* (*TGF-β* signaling and matrisome;
206 *CST1+, TGFb1+*), which were previously divided into two major subtypes: iCAF and myCAF (**Fig.
207 3b**). Among these five subtypes, *ecm-myCAF* and *TGFβ-myCAF* are known to correlate with
208 immunosuppressive environments and are enriched in tumors with high regulatory T lymphocytes
209 (Tregs) and depleted CD8+ lymphocytes. Additionally, these subtypes are associated with
210 primary immunotherapy resistance in melanoma and lung cancer[49].

211

212 The CAF-S4 population expressed pericyte markers (*RGS5+, CSPG4+,* and *PDGFRA*+), *CD248
213 (endosialin)*, and *EPAS1 (HIF2-α),* that this particular CAF subtype is vessel-associated, with
214 hypoxia potentially contributing to invasion and metastasis, as has been shown in another study
215 [48]. CAF-S4 clustered into the immature phenotype (*RGS5+, PDGFRB+,* and *CD36+)* and the
216 differentiated myogenic subtype (*TAGLN*+ and *MYH11+*)(**Fig. 3c and Supplementary Table
217 4**)[50].

218

219 Given the correlation between CMS4 and fibroblast infiltration, we next sought to test the
220 existence of CAF-S1 and CAF-S4 signatures in bulk transcriptomic data and their association with
221 clinical outcomes[15]. To this end, we interrogated and carried out a meta-analysis of eight

222 colorectal cancer transcriptomic datasets comprising 1,584 samples. We detected a strong and
223 positive correlation between specific gene expressions characterizing each CAF subtype in CRC,
224 pancreatic cancer, and non-small cell lung cancer (NSCLC) (**Fig. 4**) (see methods for datasets).
225 The gene signatures were specific to each CAF-S1 and CAF-S4, thus confirming their existence
226 in colorectal cancers and other tumor types.
227
228 We found high CAF-S1 and CAF-S4 signatures associated with poor median overall-survival and
229 early cancer recurrence (HR>1, $p<0.05$), irrespective of CMS subtypes in three independent CRC
230 datasets (**Supplementary Fig 7**). Additionally, CAF signatures stratified the CMS4 subtype into
231 high- and low-risk overall survival in all datasets, thus identifying additional heterogeneity and
232 providing prognostication in this aggressive patient subgroup (**Fig. 4b-d**). Here, using scRNA-
233 seq, we show for the first time that high CAF infiltration in CRC is associated with poor prognosis
234 across all molecular subtypes, and which further stratifies the CMS4 subgroup into high and low-
235 risk clinical phenotypes in CRC cohorts.
236
237 **Single-cell RNA sequencing reveals heterogeneity beyond Consensus Molecular Subtypes**
238 **in colorectal cancers and offers therapeutic opportunities.**
239
240 The lack of association between tumor epithelia and CMS classification, as well as the survival
241 differences between high- and low-risk CAF signatures across CRC molecular subtypes suggest
242 CRCs are much more heterogeneous than the traditional classification systems have indicated
243 (e.g. those systems defined by somatic alterations, epigenomic features, and bulk gene
244 expression data)[13,14,51,52].
245
246 Among these the widely adopted CMS classification, which reflects both the malignant cell
247 phenotypes and the TME, classified CRC into CMS1 (MSI immune), CMS2 (canonical), CMS3

248 (metabolic), and CMS4 (mesenchymal) subtypes based on bulk transcriptomic signatures[15]. To
249 test our hypothesis, we estimated every cell type fraction using single-cell data from eight pooled
250 datasets (>1,500 samples) with a machine-learning algorithm, CIBERSORTx[53]. When we
251 compared epithelial, immune, and stromal cell populations among the CMS subtypes, we did not
252 detect a distinct pattern of tumor, immune, or stromal cell signatures across the different CMS
253 subtypes. Each CMS subtype was enriched in these cell types in varying proportions but without
254 a clear distinctions between the four subtypes, suggesting a lack of clear separation among the
255 CMS subsets at the single-cell resolution (**Fig. 5a-b**). Upon analysis of four independent bulk RNA
256 datasets, there was significant discordance in terms of cell phenotype enrichment with respect
257 to each CMS subtype across the datasets except CMS4 which had predominant stromal
258 enrichment (**Supplementary Figs. 8-11)**[54]. These discordant results could be due to intra-patient
259 CMS heterogeneity, intratumoral variation in tumor purity, stromal and immune cell infiltration,
260 and CMS's inability to address tumor/TME-to-tumor/TME variability, among others[16]. Thus, novel
261 approaches that consider these factors are required to stratify patients for optimal biomarker and
262 therapeutic development.
263
264 Based on the above findings, we postulate that CRC is more accurately represented in a
265 transcriptomic continuum previously proposed by Ma et al.[55]. The authors analyzed bulk
266 transcriptomic data using a novel computational framework in which denovo, unsupervised
267 clustering methods (k-medoid, non-negative-matrix factorization, and consensus clustering)
268 demonstrated the existence of CRC in a transcriptomic continuum[56–58]. They further carried out
269 principal component analysis and robustly validated two principal components, PC Cluster
270 Subtype Scores 1 and 2 (PCSS1 and PCSS2, respectively). We reasoned that using single-cell
271 data could deepen our understanding of how each cell phenotype contributes to the CRC tumor
272 microenvironment using continuous scores that inform CRC diversity beyond binning CRC into
273 traditional classifications.

274

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

281

282 To test continuous scores reproducibility, we analyzed four bulk transcriptomic datasets
283 separately; we found that transcriptional shifts were reproducible across datasets for all major cell
284 types (**Supplementary Figs. 8-11)**. The continuous scores showed no reliability in classifying
285 CRC into immune–stromal rich (CMS1/CMS4) or immune-stromal desert (CMS2/CMS3) subtypes
286 as proposed previously [21]. Thus, confirming continuous scores rather than discrete subtypes may
287 improve classifying CRC tumors and may explain tumor-to-tumor variability, tumor/TME-to-
288 tumor/TME and TME-to-TME variabilities [55]. Of note, CAF-S1 exhibited high PCSS1 and PCSS2
289 scores across independent datasets, correlating with the CMS4 subtype. Thus, our analysis
290 identified CAF-S1 as a cell of origin for biological heterogeneity in CMS4 subtypes associated
291 with poor prognosis (**Supplementary table 5**).

292

293 **DISCUSSION**

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

300 sample set of 1,182 high-quality fibroblasts, we identify the CAF-S1 subtype to be the cell of origin
301 associated with poor prognosis across all CRC CMS subtypes and not just CMS4 [59]. Further, we
302 developed a novel signature of CAF infiltration and demonstrate that CMS4 can be stratified into
303 risk groups associated with good or poor median overall survival. These findings are significant
304 since the CMS4 subtype, is primarily stromal-driven and is enriched in more than 40% of
305 metastatic CRC samples from patients with worse outcomes[22]. We also identified CAF-S1
306 subtypes associated with certain biological functions in other cancers, including the *ecm-myCAF*
307 and *TGFβ-myCAF* subtypes (responsible for immunotherapy resistance in NSCLC and
308 melanoma), and CAF-S4s known to play a role in inducing cancer cell invasion[49,50]. Thus,
309 targeting CAFs to remodel the tumor microenvironment may lead to improved and much-needed
310 therapeutic development for metastatic CRC patients[22,23].
311
312 Targeting of CAFs in solid tumors is being explored in multiple clinical trials with variable results[60–
313 62]. Such studies likely failed to address CAF heterogeneity and their complex interactions with the
314 other cells of TME. Our study suggests that CRC may be intricately entwined with the stroma,
315 and therefore may be amenable to stromal targeted combinatoric approaches, including
316 monoclonal antibodies that abrogate CAF-S1 function. In future studies, the treatment of CRC
317 patients should involve stroma targeted therapies and take the above aspects into
318 consideration[60,63]. The scRNA-seq or bulk-RNA-seq signatures corresponding to CAF-S1 and
319 CAF-S4 may serve as suitable biomarkers for tumors that are reliant on this axis. Immunotherapy
320 responses in MSS CRC, which comprise almost 95% of metastatic CRC, are lacking; CAF
321 subpopulations within the TME may be suppressing immune responses in these tumors [64]. Based
322 on our analysis, we speculate that targeting CAF-derived chemokines and cytokines via
323 biospecific antibodies, vaccines, or even cell-based therapies, may enhance current checkpoint
324 blockade strategies[60]. Functional validation and clinical studies will be required to confirm the
325 clinical utility of targeting these CAF populations in CRC.

326

327 More importantly, our study's single-cell resolution enables us to investigate whether tumor cell
328 transcriptomes, and by extension, biological phenotypes, are the primary determinant of CMS
329 classification. Based on our findings, it appears that bulk analysis may have been confounded by
330 varying tumor microenvironment population enrichment, and that tumor cells within each patient
331 do not segregate into static phenotypes but rather exhibit considerable plasticity. In contrast, the
332 single-cell analysis uncovered the complex and mixed cellular-phenotypes among each cell
333 specific subpopulation, which projected in a transcriptomic continuum across CMS subtypes.
334 These findings were further supported by scRNA-seq CMS classification analysis that assigned
335 each CRC sample to multiple CMS subtypes thereby suggesting CMS heterogeneity in each CRC
336 tumor[21,65]. These findings may also explain why CMS-defined populations of tumors have not
337 been readily observed in transcriptomic data from independent CRC cohorts[22,54]. Our data
338 indicate that attempts to divide CRC phenotypes into the current discrete subtypes may
339 undermine optimal patient stratification in the clinical trial setting. Intriguingly, by applying two
340 independent algorithms, we demonstrate that CRC tumors and their ecosystems exist in a
341 transcriptomic continuum and not only show tumor-to-tumor variability (as proposed by Ma et al.)
342 but also demonstrate tumor/TME-to-tumor/TME transcriptional variability at the single-cell
343 resolution[55]. The continuous scores are reproducible across transcriptomic datasets, thus
344 allowing robust identification of patient subtypes. This may help to optimize CRC treatment in
345 future studies.

346

347 In conclusion, our study lends strong support to the tumor biology models proposed by Ma et al.
348 (and other groups) and represents a conceptual shift in our understanding of CRC pathogenesis,
349 clinical management, and therapeutic development. Future studies will need to consider tumor-
350 TME to tumor-TME heterogeneity which will be critical for optimizing biomarkers and treatment
351 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*.'

385 **METHODS**

386

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

395

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

407

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

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

437 using the Seurat *FindAllMarkers* function. We used positively expressed genes with an average
438 expression of >/= 2-fold higher in that subcluster than the average expression in the rest of the
439 other subclusters. We used known marker genes, which have the highest fold expression in that
440 cluster with respect to the other clusters. We also utilized SingleR ((v0.99.10, R Package), which
441 leverage large transcriptomic datasets of well-annotated cell types and manual annotation for
442 cell-type identification[32,73–75]. Depending on the presence of known marker genes the clusters
443 were grouped as: epithelial cells (*EPCAM, KRT8,* and *KRT18*), fibroblasts (*COL1A1, DCN,*
444 *COL1A2,* and *C1R*), endothelial cells (*CLDN5, FLT1, CDH5,* and *RAMP2*), myeloid cells (*LYZ,*
445 *MARCO, CD68,* and *FCGR3A*), CD4 T cells (*CD4*), CD8 T cells (*CD8A* and *CD8B*), and B cells
446 (*MZB1*), [32,48,73,76–80]. The cells were eventually assembled into DGE matrices within each
447 compartment, containing all six cell types.
448
449 ***Major-cell type subclustering and data visualization.*** Each major cell type, including epithelial
450 cells, endothelial cells, T cells, B cells, myeloid cells, and fibroblasts was reclustered and
451 reanalyzed to study each compartment at a higher resolution to detect granular cellular
452 heterogeneity in CRC. Clustree (v0.4.3) and manual review were utilized for optimal cluster
453 detection. For cell annotation of each cell type, we utilized published literature gene expression
454 signatures and manual review of differential genes among clusters. Additionally, we again utilized
455 SingleR (v0.99.10, R Package) for unbiased cell annotation. Interestingly, reclustering of major
456 compartments individually also detected clusters expressing hybrid markers as well as cell
457 clusters expressing markers from distinct lineages (such as T cell clusters expressing B cells);
458 these were removed and excluded for further analysis. We utilized UMAP for visualization
459 purposes. For validation, we analyzed 65,362 cells from 23 patients and applied the same quality
460 control metrics as outlined above, retaining 31,383 high-quality single cells for further analysis[30].
461 These high-quality cells were analyzed utilizing the same pipelines and parameters as that for our
462 primary cohort (**Supplementary Figs. 4-5 and 12-13**).

463 The InferCNV (v1.2.1) package was used with default paramets to identify the evidence for
464 somatic large-scale chromosomal copy number alteration in epithelial cells (*EPCAM+, KRT8+,*
465 *KRT18+*)[81]. Non-malignant epithelial cells were used as the control group.
466
467 ***Trajectory analysis.*** We used Monocle v.2 (v2.14.0), a reverse graph embedding method to
468 reconstruct single-cell trajectories in tumor and non-malignant epithelium[82]. In brief, we used UMI
469 count matrices and the negbinomial.size() parameter to create a *CellDataSet* object in the default
470 setting. We grouped projected cells on UMAP in default settings for visualization of monocle
471 results. We defined the cumulative duration of the trajectory to show the average amount of
472 transcriptional transition that a cell undergoes as it passes from the starting state to the end state.
473 The cells were also ordered in pseudotime to explain the transition of cells from one state to
474 another.
475
476 ***Pathway- Gene set variation analysis (GSVA).*** Pathway analysis was performed on the 50
477 hallmark gene sets downloaded from *Molecular Signatures Database (v7.2)*. We used GSVA
478 (v1.34.0), a non-parametric, unsupervised method to estimate the gene set variations and
479 evaluation of pathway enrichment, and pathway scores were calculated for each cell using
480 standard settings [34,39].
481
482 ***DNA and bulk RNA library construction.*** DNA and bulk RNA sequencing was performed as
483 previously described[83]. One hundred nanograms of DNA from each tumor was mechanically
484 sheared to an average size of 200 bp. Using the *KAPA Hyper Prep Pack*, DNA libraries were
485 packed, hybridized into the *xT probe* package, and amplified with the *KAPA HiFi HotStart*
486 *ReadyMix*. For uniformity, each sample needed to have 95% of all targeted base pairs sequenced
487 to a minimum depth of 300x. One hundred nanograms of RNA per tumor sample was heat
488 fragmented to a mean size of 200 base pairs in the presence of magnesium. Using random

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

515 listed as MSE. If there was a <30% chance of MSI-H status, the sample was called MSS.
516 Additionally, IHC results were used to classify tumors into MSS or MSI molecular subtypes. Both
517 of these modalities were concordant and produced the same results.
518
519 ***Bulk RNA-seq and microarray analysis.*** We downloaded and pooled eight colorectal gene
520 expression datasets (GSE13067[85], GSE13294[85], GSE14333[86], GSE17536[87], GSE20916[88],
521 GSE33113[89], GSE35896[90], and GSE39582[14]), a pancreatic cancer dataset (GSE62165[91]) and
522 a non-small cell lung cancer dataset (GSE33532[92]) to validate our findings from the single cell
523 compartments by deconvoluting the bulk gene expression profiles into pseudo single-cell
524 resolutions. We used Affy (v1.64.0) for the data analysis and for exploration of Affymetrix
525 oligonucleotide array probe level data[93]. Batch correction was carried out using the
526 removeBatchEffect (v3.42.2) function of the LIMMA program and CMScaller for the CMS
527 classification (see below)[94]. Three datasets (GSE17536[87], GSE33113[89], and GSE39582[14]) were
528 utilized for clinical outcome analysis[94,95].
529
530 ***Correlation patterns in bulk gene expressions for CAF compartments.*** To identify the top
531 correlated CAF-marker genes within the combined eight and individual four bulk gene expression
532 sets, we first transformed the bulk gene expression sets with $\log_2$ transformation. Next, marker
533 genes with an average $\log_2$ FC>/= 0.5 and $p<0.05$ obtained from the SC analysis of CAF-S1 and
534 CAF-S4 compartments were separately intersected with the bulk gene expression sets. Genes
535 with an average Spearman correlation score greater than 0.8 were kept as the CAF signatures
536 within the bulk gene expression. Heatmaps illustrating the correlation patterns within and between
537 the CAF compartments were prepared with the heatmap.2 function from ggplot package (v3.1.1)
538 utilizing the Pearson correlation coefficient. Heatmaps illustrating the correlation patterns within
539 and between the CAF compartments were prepared using the ggplot package (v3.1.1) utilizing
540 the Pearson correlation coefficient[96].

541

542 The Cox proportional hazard regression model was used to examine the significance of 20 cell
543 types from scRNA-seq in bulk expression data. Each cell type's marker genes with an average
544 logFC>1 and adjusted P<0.05 were intersected with the bulk expression datasets separately. We
545 only kept the marker genes with a high correlation with each other in bulk, which provides an
546 average correlation score of > 0.8. The average bulk expression of each cell type's remaining
547 marker genes was calculated and used in the hazard regression model as the representative of
548 this cell type. For analysis of relationships with patient outcome, univariate models were
549 calculated using Cox proportional hazard regression (coxph function from survival R package)[97].

550

551 ***Deconvoluting public bulk gene expression profiles into pseudo single-cell expressions.***
552 We used CIBERSORTx v1 to estimate composition of various cell populations in pooled eight
553 microarray datasets[53]. Signature gene matrices were created using the expression profiles of
554 49,859 cells as the reference single cell profile. We ran the 'hires' module with default parameters
555 except for the 'rmbatchBmode,' and the bulk-mode batch correction argument was set to true.
556 After the deconvolution process, we normalized the gene expressions according to the cell
557 fractions in each sample and calculated each gene's Z-transformed expression values. The
558 average normalized expression of each cell type across all samples was plotted with the
559 heatmap.3 R function of the GMD package (v0.3.3)[98]. A signature matrix highlighting marker
560 genes of the different cell types was prepared with a heatmap.2 R function of ggplot (v3.1.1). We
561 also applied the same parameters to deconvolute GSE14333[86], GSE17536[87], GSE33113[89], and
562 GSE39582[14] datasets individually.

563

564 ***Consensus molecular subtyping of colorectal cancer (CMS Classification).*** We used R
565 package CMScaller(v0.9.2), a nearest template prediction (NTP) algorithm, for the classification
566 of gene expression datasets[95]. We set the permutation number to 1000 to predict the CMS classes

567 of the samples in the GEO datasets with a p-value < 0.05. We ran CMScaller with default
568 parameters.
569
570 ***Continuous subtype discovery using scRNA-seq analysis.*** Bulk mRNA expression profiles of
571 the combined and batch adjusted eight GEO datasets (GSE13067[85], GSE13294[85], GSE14333[86],
572 GSE17536[87], GSE20916[88], GSE33113[89], GSE35896[90], and GSE39582[14]), composed of 1584
573 samples in total, were deconvoluted into the pseudo single-cell expression profiles via
574 CIBERSORTx utilizing the expression data consisting of 2o different cell types from our scRNA-
575 seq dataset[53]. We transformed the deconvoluted expression matrix with log2 transformation. The
576 principal components cluster subtype scores (PCSSs) of the CMS subtypes among the 1584
577 samples, were determined separately for each cell type using an algorithm published by Ma et
578 al[55]. To obtain the PCSSs, the average loading vectors were used. The results obtained for 20
579 cell types were projected on the first two PCSSs (PCSS1 and PCSS2) as they were validated by
580 Ma et al. in their analysis using 18 datasets. We also analyzed four datasets (GSE14333[86],
581 GSE17536[87], GSE33113[89], and GSE39582[14]) to independently confirm reproducibility of
582 continuous scores.
583
584 ***Statistics and reproducibility.*** All statistical analyses and graphs were created in R (v3.6.3) and
585 using a Python-based computational analysis tool. Schematic representations were made using
586 the Inkscape (https://inkscape.org/) software. Dim plots, bar plots and box plots were generated
587 using the dittoSeq (v1.1.7) package with default parameters[99]. Violin plots were generated using
588 the patchwork (v1.1.0) package and ggplot2 (v3.3.2) package in R with default parameters.
589 Heatmaps were generated using Morpheus.R with default parameters[100,101]. ANOVA and pair-
590 wised t-tests for the CMS classes across the deconvoluted expression profiles were performed in
591 R using the ggpubr R (v0.4.0) package[102]. The Box and Whisker plots were generated using the
592 boxplot function of the R base package at default parameters. The mean of the log_2 transformed

593 deconvoluted expression value of the samples in each CMS group was demonstrated with a
594 horizontal straight line within each box. The length of a boxplot corresponds to the interquartile
595 range (IQR), which is defined as the range between the first and third quartiles (Q1 and Q3),
596 whereas the whiskers are the upper and lower extreme values of the data (either data's extremum
597 values, or the Q3+1.5*IQR and Q1-1.5*IQR values, whichever was less extreme).
598
599 ***Survival analysis.*** Survival curves were obtained according to the Kaplan-Meier method survfit
600 (v3.2-7), and differences between survival distributions were assessed by Log-rank test. The
601 patients were divided into two groups (high/poor and low/good risk) according to their median
602 expression values (survminer (v0.4.8)). The surv_cutpoint function uses the maximally selected
603 rank statistics and implements standard methods for the approximation of the null distribution of
604 maximally selected rank statistics (maxstat (v0.7-25).
605
606 The proportional hazard assumption was tested to examine the fit of the model for survival of the
607 samples in four GEO datasets (GSE14333[86], GSE17536[87], GSE33113[89], and GSE39582[14]) with
608 respect to the deconvoluted bulk mRNA expressions. For analysis of the relationships with patient
609 outcome, multivariate models were calculated using the Cox proportional hazard regression
610 (coxph survival R package)[97].
611
612
613
614
615
616
617
618

619 **REFERENCES**
620

621 1. Cancer of the Colon and Rectum - Cancer Stat Facts. *SEER*
622 https://seer.cancer.gov/statfacts/html/colorect.html.
623 2. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut*
624 **66**, 683–691 (2017).
625 3. Osterman, E. *et al.* Recurrence risk after radical colorectal cancer surgery—less than
626 before, but how high is it? *Cancers* **12**, 3308 (2020).
627 4. Molinari, C. *et al.* Heterogeneity in colorectal cancer: a challenge for personalized medicine?
628 *Int J Mol Sci* **19**, (2018).
629 5. Xie, Y.-H., Chen, Y.-X. & Fang, J.-Y. Comprehensive review of targeted therapy for
630 colorectal cancer. *Signal Transduction and Targeted Therapy* **5**, 1–30 (2020).
631 6. Budinska, E. *et al.* Gene expression patterns unveil a new level of molecular heterogeneity
632 in colorectal cancer. *J Pathol* **231**, 63–76 (2013).
633 7. Deschoolmeester, V. *et al.* Tumor infiltrating lymphocytes: an intriguing player in the survival
634 of colorectal cancer patients. *BMC Immunol* **11**, 19 (2010).
635 8. Lee, W.-S., Park, S., Lee, W. Y., Yun, S. H. & Chun, H.-K. Clinical impact of tumor-
636 infiltrating lymphocytes for survival in stage II colon cancer. *Cancer* **116**, 5188–5199 (2010).
637 9. Perez, E. A. *et al.* Association of stromal tumor-infiltrating lymphocytes with recurrence-free
638 Survival in the N9831 adjuvant trial in patients with early-stage HER2-positive breast
639 cancer. *JAMA Oncology* **2**, 56–64 (2016).
640 10. Nearchou, I. P. *et al.* Spatial immune profiling of the colorectal tumor microenvironment
641 predicts good outcome in stage II patients. *npj Digital Medicine* **3**, 1–10 (2020).
642 11. Pagès, F. *et al.* International validation of the consensus Immunoscore for the classification
643 of colon cancer: a prognostic and accuracy study. *The Lancet* **391**, 2128–2139 (2018).
644 12. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human
645 colon and rectal cancer. *Nature* **487**, 330–337 (2012).
646 13. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular
647 phenotype and responses to therapy. *Nature Medicine* **19**, 619–625 (2013).
648 14. Marisa, L. *et al.* Gene Expression Classification of Colon Cancer into Molecular Subtypes:
649 Characterization, Validation, and Prognostic Value. *PLOS Medicine* **10**, e1001453 (2013).
650 15. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine*
651 **21**, 1350–1356 (2015).
652 16. Roepman, P. *et al.* Colorectal cancer intrinsic subtypes predict chemotherapy benefit,
653 deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of*
654 *Cancer* **134**, 552–562 (2014).
655 17. Berg, I. van den *et al.* Improving clinical management of colon cancer through
656 CONNECTION, a nation-wide colon cancer registry and stratification effort (CONNECTION
657 II trial): rationale and protocol of a single arm intervention study. *BMC Cancer* **20**, 1–8
658 (2020).
659 18. Lenz, H.-J. *et al.* Impact of Consensus Molecular Subtype on Survival in Patients With
660 Metastatic Colorectal Cancer: Results From CALGB/SWOG 80405 (Alliance). *J Clin Oncol*
661 **37**, 1876–1885 (2019).
662 19. Stintzing, S. *et al.* Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and
663 first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306)
664 trial. *Ann Oncol* **30**, 1796–1803 (2019).
665 20. Mooi, J. K. *et al.* The prognostic impact of consensus molecular subtypes (CMS) and its
666 predictive effects for bevacizumab benefit in metastatic colorectal cancer: molecular
667 analysis of the AGITG MAX clinical trial. *Ann Oncol* **29**, 2240–2246 (2018).
668 21. Sveen, A., Cremolini, C. & Dienstmann, R. Predictive modeling in colorectal cancer: time to
669 move beyond consensus molecular subtypes. *Annals of Oncology* **30**, 1682–1685 (2019).

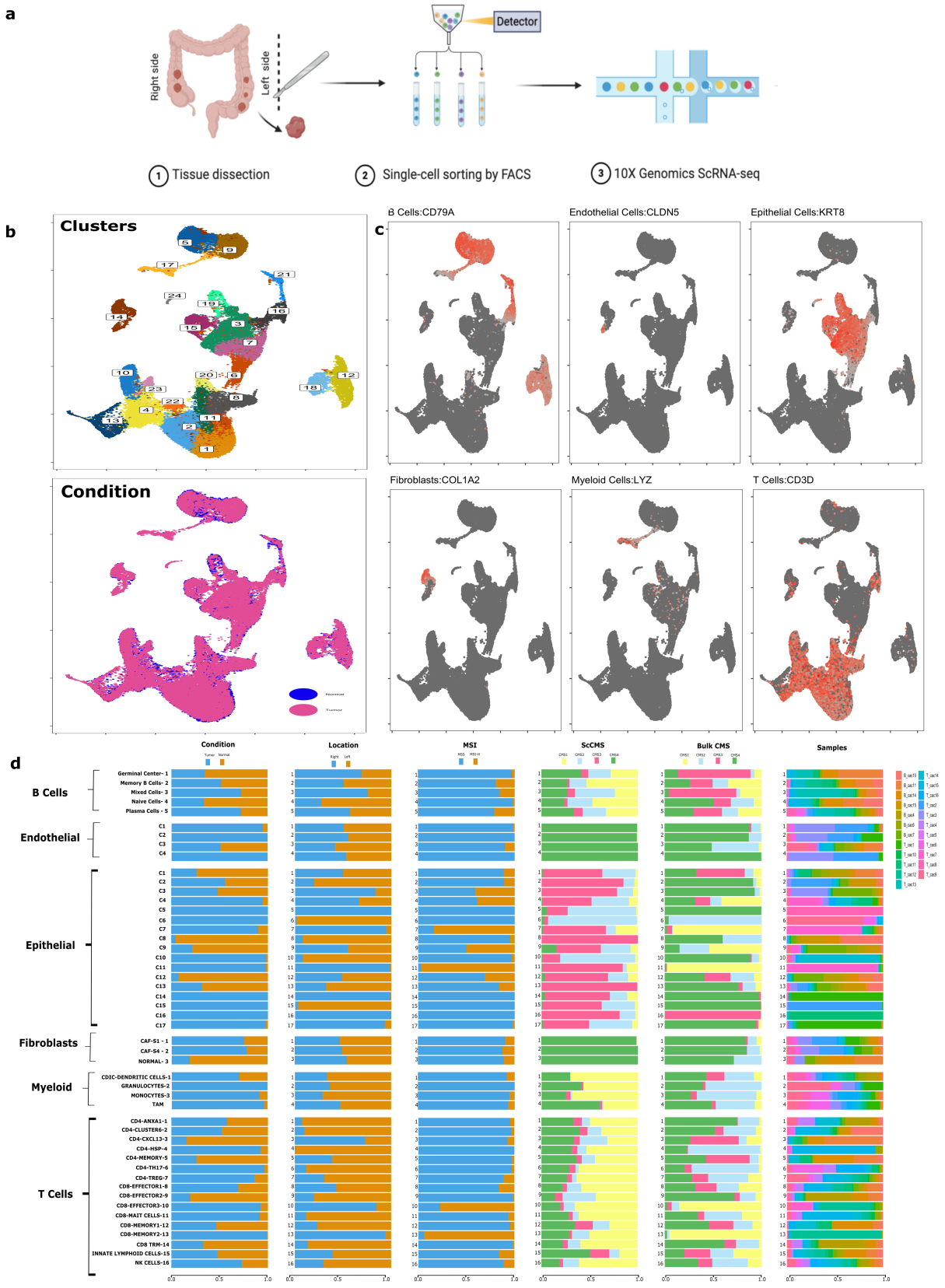
670 22. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters—
671 consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals*
672 *of Oncology* **30**, 520–527 (2019).
673 23. Khambata-Ford, S. *et al.* Expression of epiregulin and amphiregulin and K-ras mutation
674 status predict disease control in metastatic colorectal cancer patients treated with
675 cetuximab. *J Clin Oncol* **25**, 3230–3237 (2007).
676 24. Aderka, D., Stintzing, S. & Heinemann, V. Explaining the unexplainable: discrepancies in
677 results from the CALGB/SWOG 80405 and FIRE-3 studies. *The Lancet Oncology* **20**, e274–
678 e283 (2019).
679 25. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at
680 multiple resolutions. *Gigascience* **7**, (2018).
681 26. Zhang, L. *et al.* Single-cell analyses inform mechanisms of myeloid-targeted therapies in
682 colon cancer. *Cell* **181**, 442-459.e29 (2020).
683 27. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal
684 cancer. *Nature* **564**, 268–272 (2018).
685 28. Corridoni, D. *et al.* Single-cell atlas of colonic CD8 + T cells in ulcerative colitis. *Nature*
686 *Medicine* **26**, 1480–1490 (2020).
687 29. Oh, D. Y. *et al.* Intratumoral CD4+ T Cells Mediate Anti-tumor Cytotoxicity in Human
688 Bladder Cancer. *Cell* **181**, 1612-1625.e13 (2020).
689 30. Lee, H.-O. *et al.* Lineage-dependent gene expression programs influence the immune
690 landscape of colorectal cancer. *Nature Genetics* **52**, 594–603 (2020).
691 31. Mills, J. C. & Sansom, O. J. Reserve stem cells: Differentiated cells reprogram to fuel repair,
692 metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci Signal* **8**, re8 (2015).
693 32. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
694 microenvironment. *Nat Med* **24**, 1277–1289 (2018).
695 33. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nature Medicine*
696 **26**, 1271–1279 (2020).
697 34. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
698 collection. *Cell Syst* **1**, 417–425 (2015).
699 35. Jamal-Hanjani, M. *et al.* Tracking the evolution of non-small-cell lung cancer. *N Engl J Med*
700 **376**, 2109–2121 (2017).
701 36. Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of
702 subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* **10**,
703 2750 (2019).
704 37. Lim, S. B. *et al.* Addressing cellular heterogeneity in tumor and circulation for refined
705 prognostication. *Proc Natl Acad Sci U S A* **116**, 17957–17962 (2019).
706 38. Wang, R. *et al.* Single-cell dissection of intratumoral heterogeneity and lineage diversity in
707 metastatic gastric adenocarcinoma. *Nat Med* **27**, 141–151 (2021).
708 39. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray
709 and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
710 40. Sathe, A. *et al.* Single-cell genomic characterization reveals the cellular reprogramming of
711 the gastric tumor microenvironment. *Clin Cancer Res* **26**, 2640–2653 (2020).
712 41. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
713 pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
714 42. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat*
715 *Methods* **14**, 979–982 (2017).
716 43. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon
717 tumors. *Nat Biotechnol* **29**, 1120–1127 (2011).
718 44. Vermeulen, L. *et al.* Single-cell cloning of colon cancer stem cells reveals a multi-lineage
719 differentiation capacity. *PNAS* **105**, 13427–13432 (2008).

720 45. Monk, M. & Holding, C. Human embryonic genes re-expressed in cancer cells. *Oncogene*
721 **20**, 8085–8091 (2001).
722 46. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–
723 767 (1990).
724 47. Costa, A. *et al.* Fibroblast heterogeneity and immunosuppressive environment in human
725 breast cancer. *Cancer Cell* **33**, 463-479.e10 (2018).
726 48. Bartoschek, M. *et al.* Spatially and functionally distinct subclasses of breast cancer-
727 associated fibroblasts revealed by single cell RNA sequencing. *Nat Commun* **9**, 5150
728 (2018).
729 49. Kieffer, Y. *et al.* Single-cell analysis reveals fibroblast clusters linked to immunotherapy
730 resistance in cancer. *Cancer Discov* **10**, 1330–1351 (2020).
731 50. Wu, S. Z. *et al.* Stromal cell diversity associated with immune evasion in human triple-
732 negative breast cancer. *The EMBO Journal* **39**, (2020).
733 51. Sadanandam, A. *et al.* Reconciliation of classification systems defining molecular subtypes
734 of colorectal cancer. *Cell Cycle* **13**, 353–357 (2014).
735 52. De Sousa E Melo, F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct
736 subtype and develops from serrated precursor lesions. *Nat Med* **19**, 614–618 (2013).
737 53. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues
738 with digital cytometry. *Nature Biotechnology* **37**, 773–782 (2019).
739 54. Dunne, P. D. *et al.* Challenging the cancer molecular stratification dogma: intratumoral
740 heterogeneity undermines consensus molecular subtypes and potential piagnostic value in
741 colorectal cancer. *Clin Cancer Res* **22**, 4095–4104 (2016).
742 55. Ma, S. *et al.* Continuity of transcriptomes among colorectal cancer subtypes based on meta-
743 analysis. *Genome Biology* **19**, 142 (2018).
744 56. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms.
745 https://link.springer.com/article/10.1007/s10852-005-9022-1.
746 57. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with
747 confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
748 58. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC*
749 *Bioinformatics* **11**, 367 (2010).
750 59. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular
751 heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708–718 (2017).
752 60. Liu, T. *et al.* Cancer-associated fibroblasts: an emerging target of anti-cancer
753 immunotherapy. *J Hematol Oncol* **12**, 86 (2019).
754 61. Barnett, R. M. & Vilar, E. Targeted therapy for cancer-associated fibroblasts: are we there
755 yet? *JNCI: Journal of the National Cancer Institute* **110**, 11–13 (2018).
756 62. Gascard, P. & Tlsty, T. D. Carcinoma-associated fibroblasts: orchestrating the composition
757 of malignancy. *Genes Dev.* **30**, 1002–1019 (2016).
758 63. Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated
759 fibroblasts. *Nat Rev Cancer* **20**, 174–186 (2020).
760 64. Roth, A. D. *et al.* Prognostic role of KRAS and BRAF in stage II and III resected colon
761 cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00
762 trial. *J Clin Oncol* **28**, 466–474 (2010).
763 65. Laurent-Puig, P. *et al.* Colon cancer molecular subtype intratumoral heterogeneity and its
764 prognostic impact: An extensive molecular analysis of the PETACC-8. *Annals of Oncology*
765 **29**, viii18 (2018).
766 66. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902.e21
767 (2019).
768 67. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet detection in single-
769 cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* **8**, 329-337.e4 (2019).

68. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome 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

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

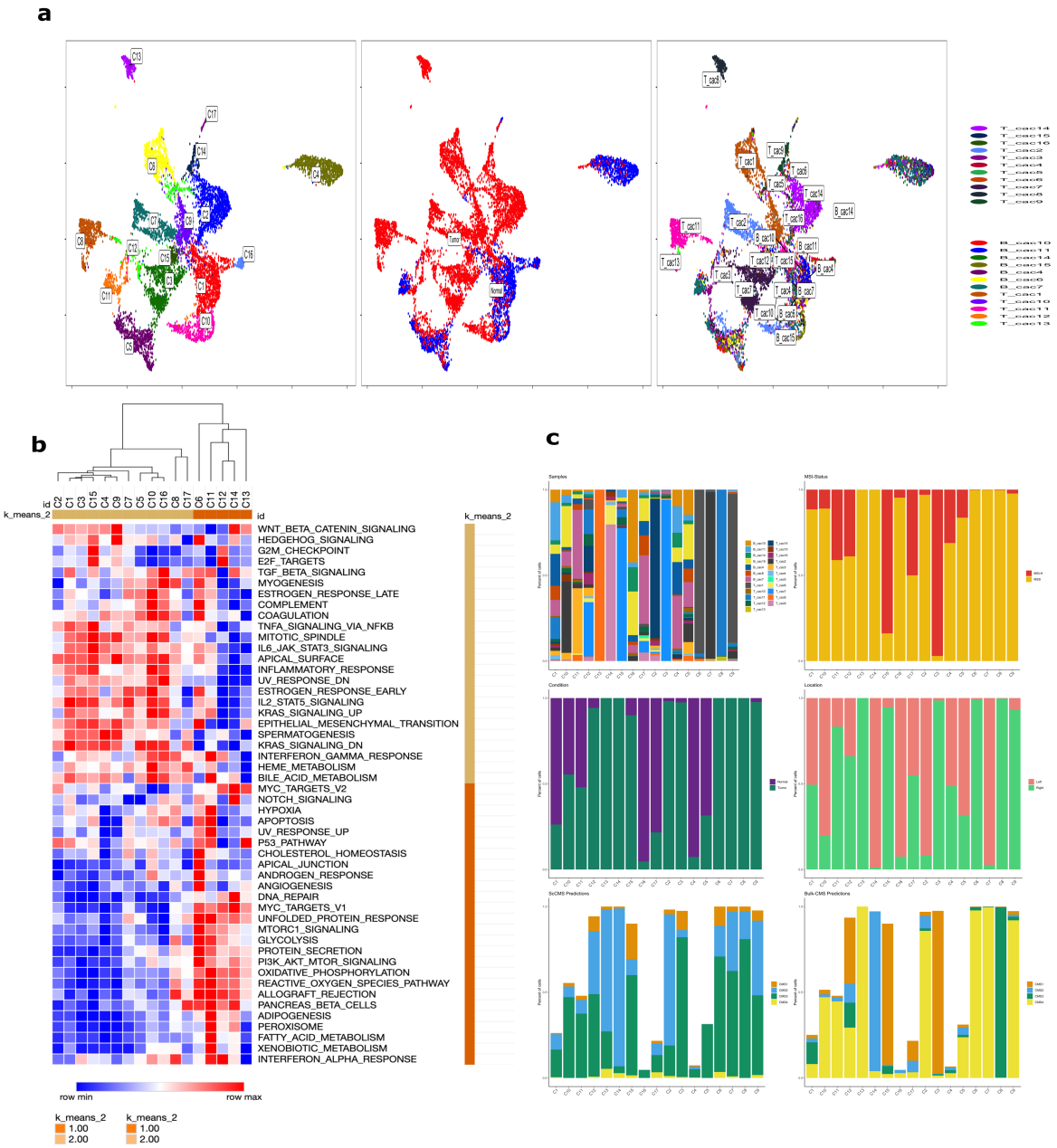
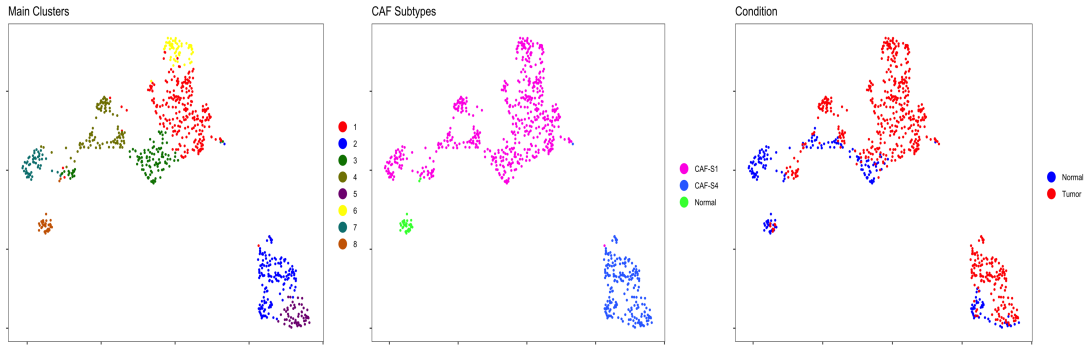
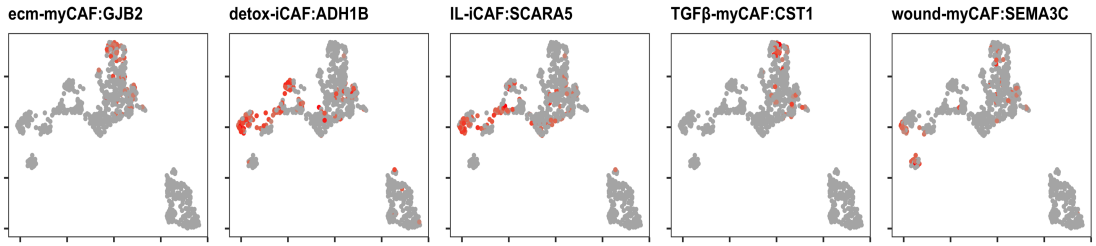

a
b
c
d
states
Location
MSI
Sc-CMS
Bulk-CMS
Component 1
Component 2
id
k_means_2
row min
row max
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
G2M_CHECKPOINT
E2F_TARGETS
TGF_BETA_SIGNALING
MYOGENESIS
ESTROGEN_RESPONSE_LATE
COMPLEMENT
COAGULATION
TNFA_SIGNALING_VIA_NFKB
MITOTIC_SPINDLE
IL6_JAK_STAT3_SIGNALING
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
Samples
MSI-Status
Condition
Location
ScCMS Predictions
Bulk-CMS Predictions

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

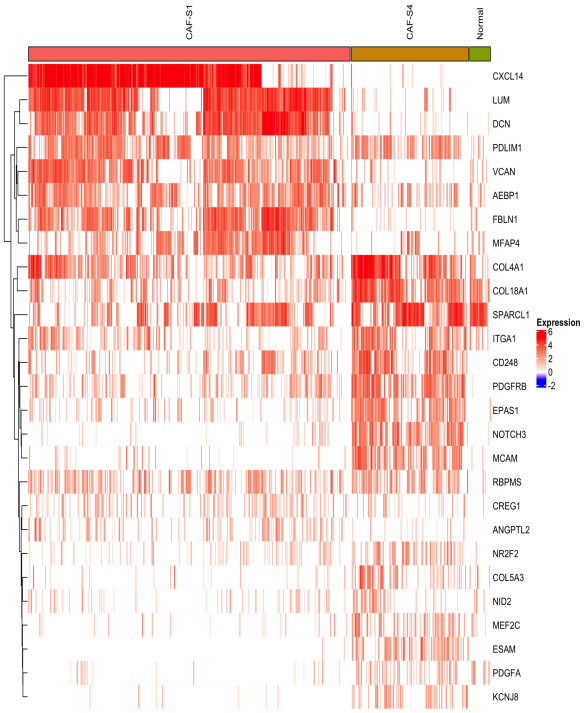
a


Main Clusters
1
2
3
4
5
6
7
8
CAF Subtypes
CAF-S1
CAF-S4
Normal
Condition
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

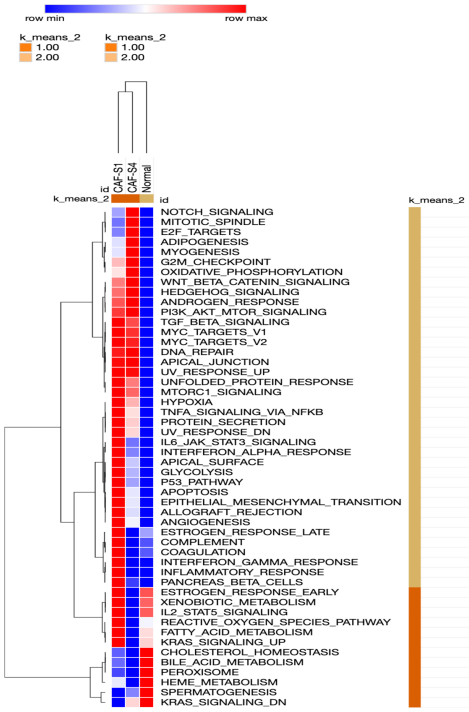

row min
row max
k_means_2
1.00
2.00
id
CAF-S1
CAF-S4
Normal
NOTCH_SIGNALING
MITOTIC_SPINDLE
E2F_TARGETS
ADIPOGENESIS
MYOGENESIS
G2M_CHECKPOINT
OXIDATIVE_PHOSPHORYLATION
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
ANDROGEN_RESPONSE
PI3K_AKT_MTOR_SIGNALING
TGF_BETA_SIGNALING
MYC_TARGETS_V1
MYC_TARGETS_V2
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

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

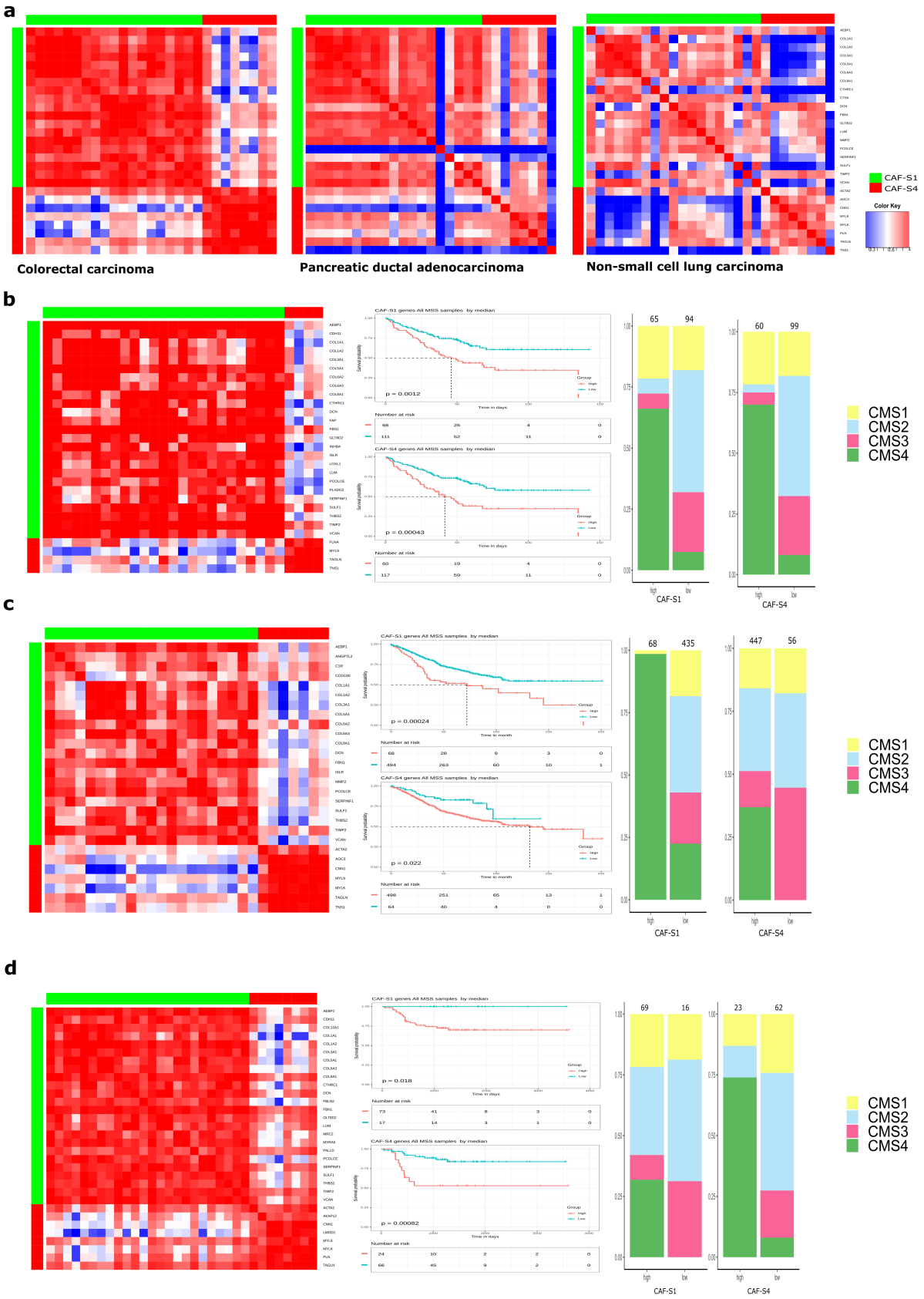

a
Colorectal carcinoma
Pancreatic ductal adenocarcinoma
Non-small cell lung carcinoma
CAF-S1
CAF-S4
Color Key
b
CAF-S1 genes All MSS samples by median
p = 0.0012
Number at risk
CAF-S4 genes All MSS samples by median
p = 0.00043
Number at risk
CMS1
CMS2
CMS3
CMS4
CAF-S1
CAF-S4
c
CAF-S1 genes All MSS samples by median
p = 0.00024
Number at risk
CAF-S4 genes All MSS samples by median
p = 0.022
Number at risk
CMS1
CMS2
CMS3
CMS4
CAF-S1
CAF-S4
d
CAF-S1 genes All MSS samples by median
p = 0.018
Number at risk
CAF-S4 genes All MSS samples by median
p = 0.00082
Number at risk
CMS1
CMS2
CMS3
CMS4
CAF-S1
CAF-S4

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= 11; 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.

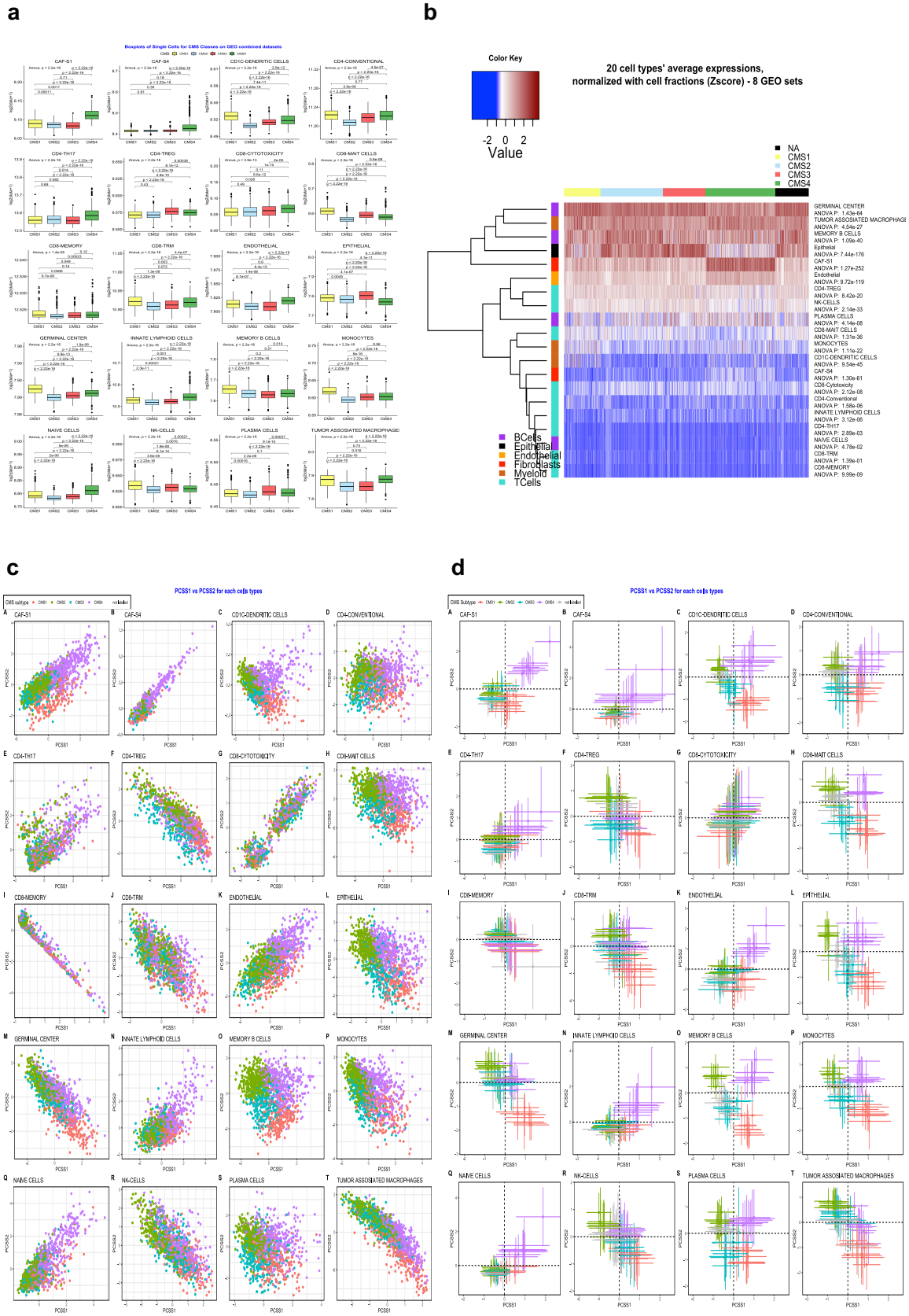
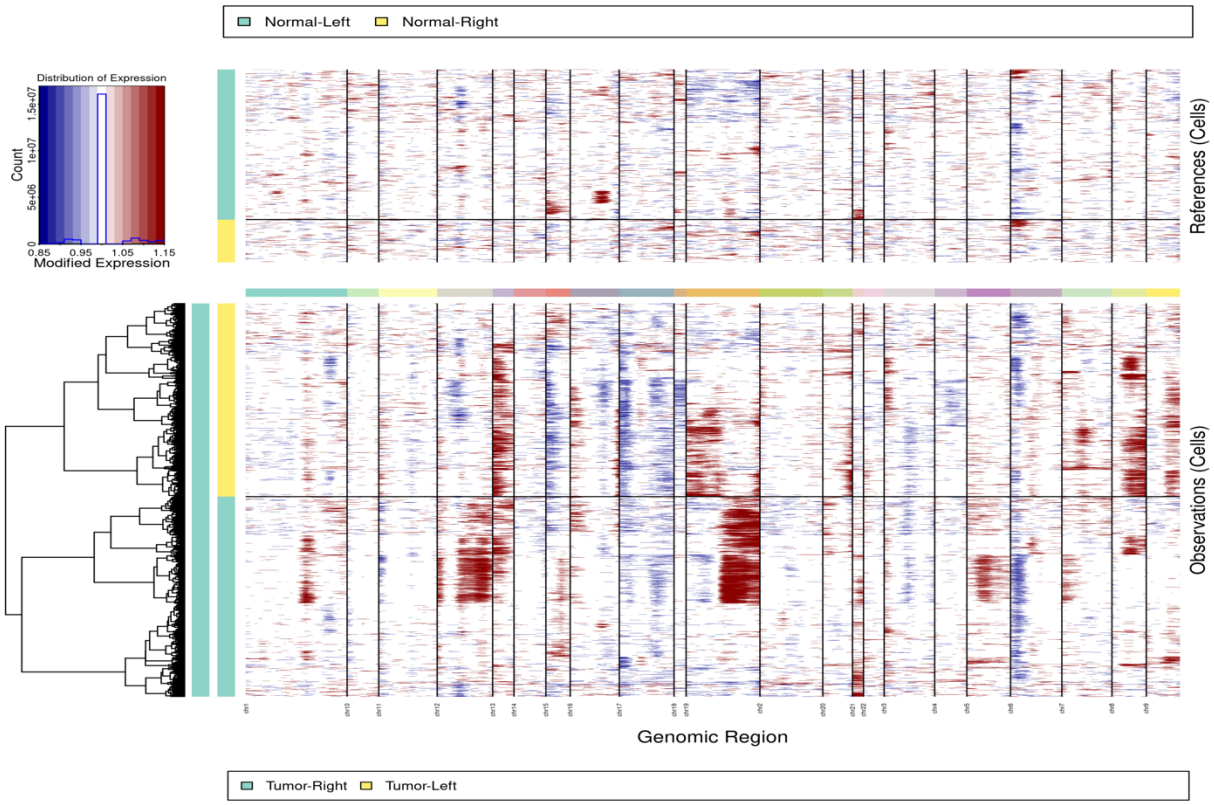
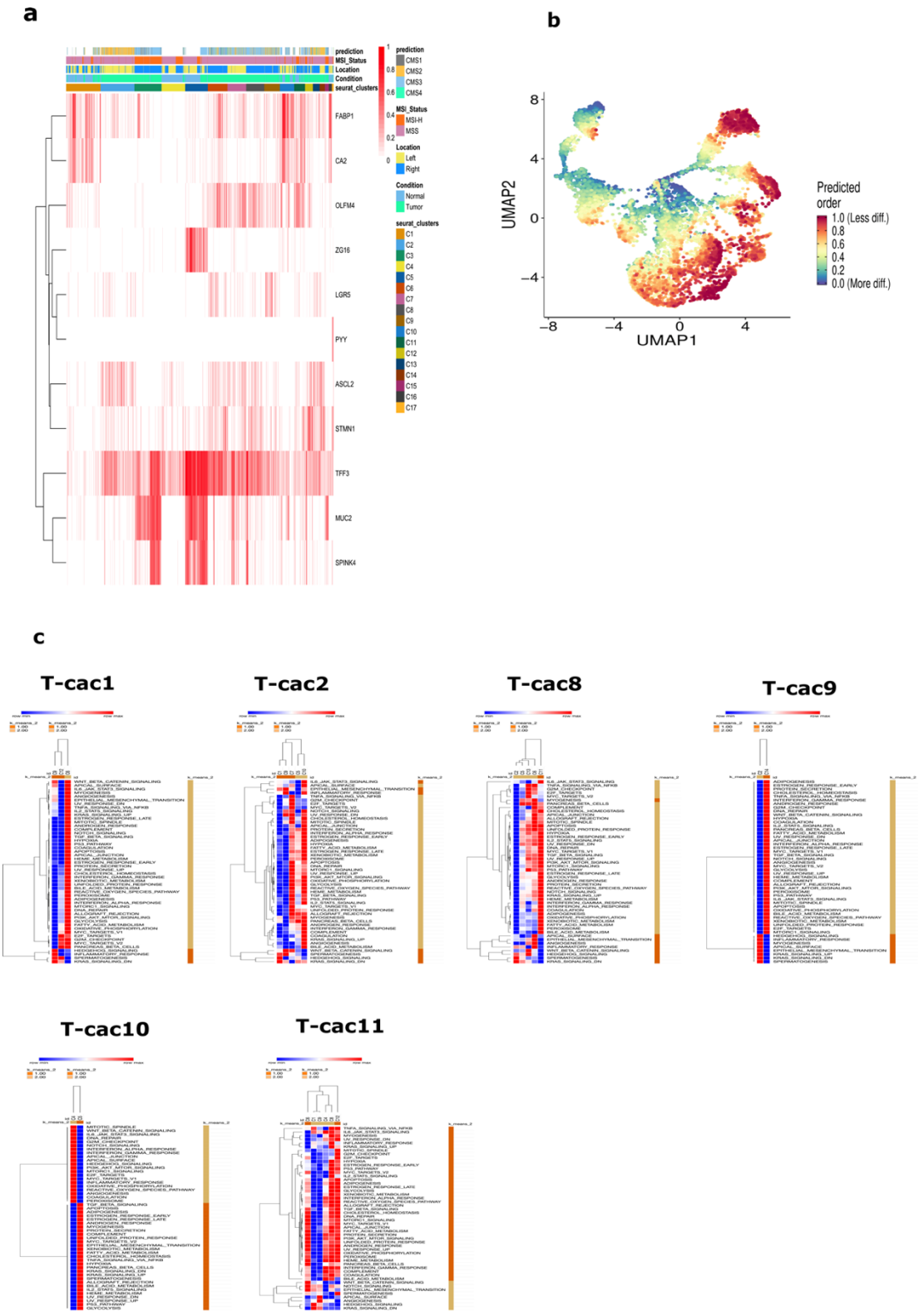

a
Boxplots of Single Cells for CMS Classes on GEO combined datasets
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TH17
CD4-TREG
CD8-CYTOTOXICITY
CD8-MAIT CELLS
CD8-MEMORY
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
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

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

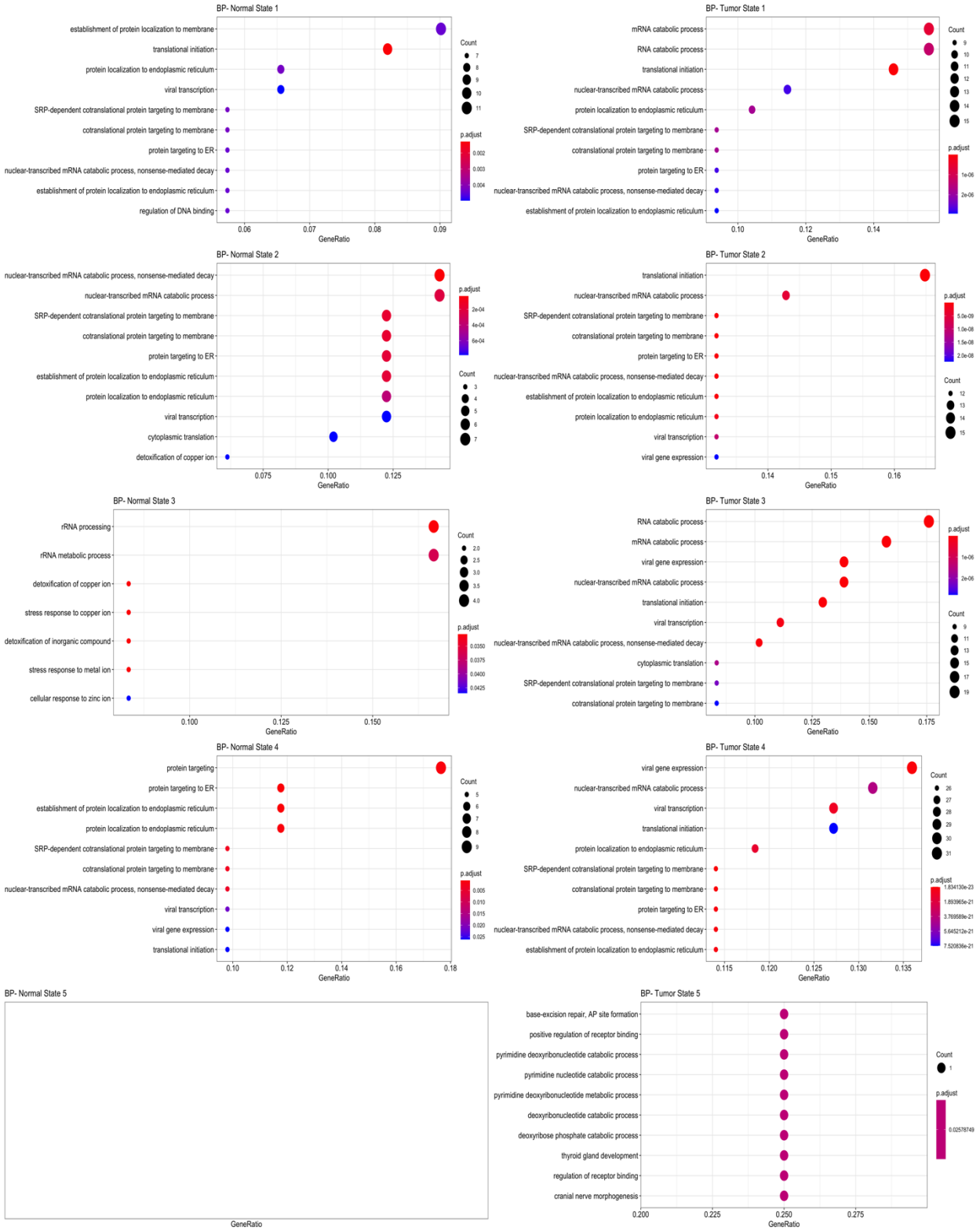


Supplementary Figure 1: Analysis of copy number variation (CNV) amongst epithelial cells. CNV analysis was conducted on the epithelial cell compartment. CNV alterations are seen in tumor-derived epithelial cells (observation). Non-malignant-derived epithelial cells were used as control (references).

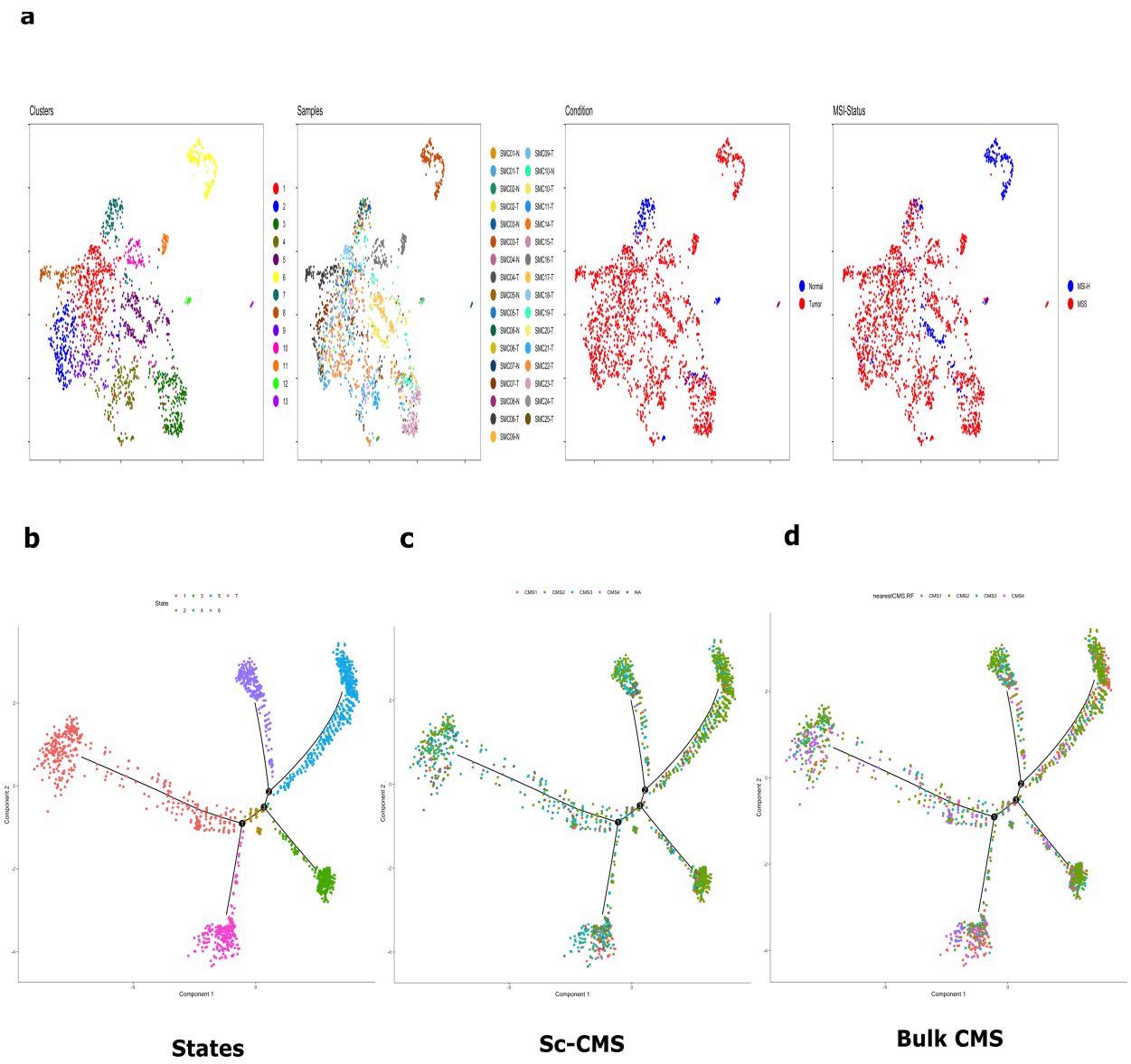


Supplementary Figure 2: Epithelial cell compartment demonstrating subclonal heterogeneity. a, Heatmap of marker gene expression for the epithelial and tumor cells. **b**,

UMAP visualization of computational analysis of differentiation status using CytoTRACE (see methods). **c**, Heatmap representation of Hallmark pathway analysis of epithelial phenotypes within six different tumor samples demonstrating subclonal, intratumoral heterogeneity.

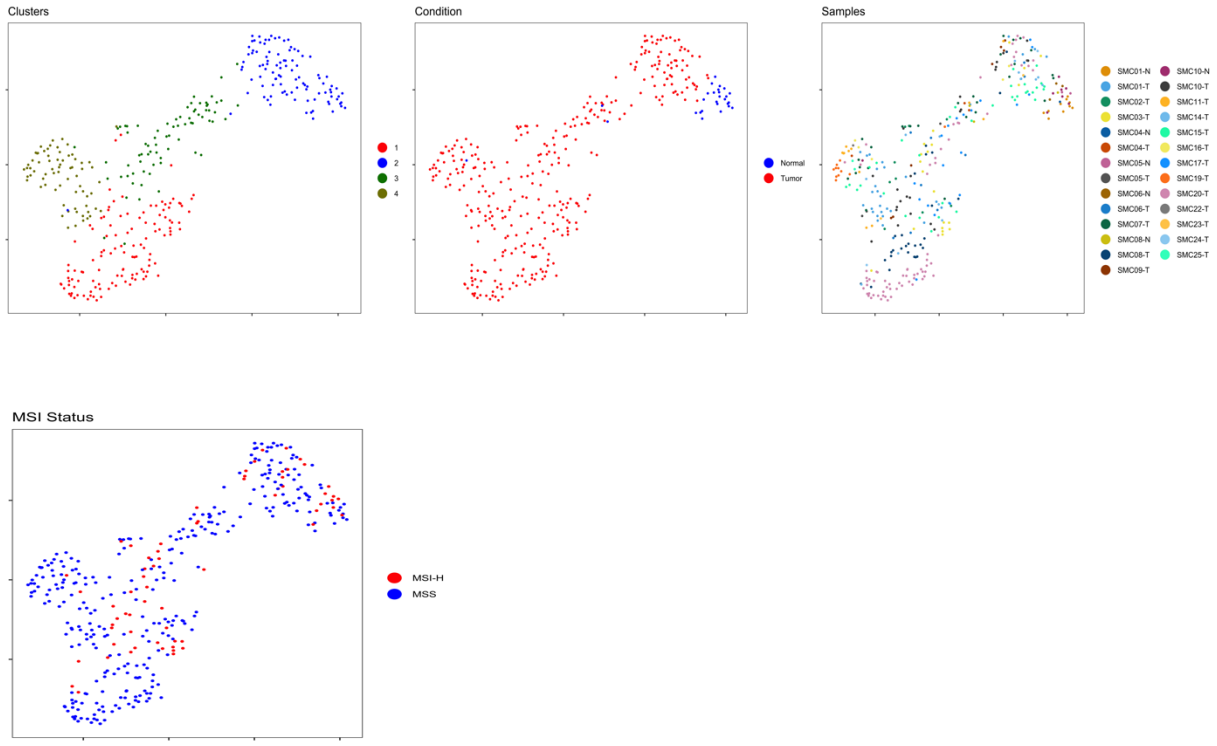

BP- Normal State 1
establishment of protein localization to membrane
translational initiation
protein localization to endoplasmic reticulum
viral transcription
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
protein targeting to ER
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
establishment of protein localization to endoplasmic reticulum
regulation of DNA binding
0.06
0.07
0.08
0.09
GeneRatio
Count
7
8
9
10
11
p.adjust
0.002
0.003
0.004
BP- Tumor State 1
mRNA catabolic process
RNA catabolic process
translational initiation
nuclear-transcribed mRNA catabolic process
protein localization to endoplasmic reticulum
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
protein targeting to ER
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
establishment of protein localization to endoplasmic reticulum
0.10
0.12
0.14
GeneRatio
Count
9
10
11
12
13
14
15
p.adjust
1e-06
2e-06
BP- Normal State 2
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
nuclear-transcribed mRNA catabolic process
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
protein targeting to ER
establishment of protein localization to endoplasmic reticulum
protein localization to endoplasmic reticulum
viral transcription
cytoplasmic translation
detoxification of copper ion
0.075
0.100
0.125
GeneRatio
p.adjust
2e-04
4e-04
6e-04
Count
3
4
5
6
7
BP- Tumor State 2
translational initiation
nuclear-transcribed mRNA catabolic process
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
protein targeting to ER
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
establishment of protein localization to endoplasmic reticulum
protein localization to endoplasmic reticulum
viral transcription
viral gene expression
0.14
0.15
0.16
GeneRatio
p.adjust
5.0e-09
1.0e-08
1.5e-08
2.0e-08
Count
12
13
14
15
BP- Normal State 3
rRNA processing
rRNA metabolic process
detoxification of copper ion
stress response to copper ion
detoxification of inorganic compound
stress response to metal ion
cellular response to zinc ion
0.100
0.125
0.150
GeneRatio
Count
2.0
2.5
3.0
3.5
4.0
p.adjust
0.0350
0.0375
0.0400
0.0425
BP- Tumor State 3
RNA catabolic process
mRNA catabolic process
viral gene expression
nuclear-transcribed mRNA catabolic process
translational initiation
viral transcription
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
cytoplasmic translation
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
0.100
0.125
0.150
0.175
GeneRatio
p.adjust
1e-06
2e-06
Count
9
11
13
15
17
19
BP- Normal State 4
protein targeting
protein targeting to ER
establishment of protein localization to endoplasmic reticulum
protein localization to endoplasmic reticulum
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
viral transcription
viral gene expression
translational initiation
0.10
0.12
0.14
0.16
0.18
GeneRatio
Count
5
6
7
8
9
p.adjust
0.005
0.010
0.015
0.020
0.025
BP- Tumor State 4
viral gene expression
nuclear-transcribed mRNA catabolic process
viral transcription
translational initiation
protein localization to endoplasmic reticulum
SRP-dependent cotranslational protein targeting to membrane
cotranslational protein targeting to membrane
protein targeting to ER
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
establishment of protein localization to endoplasmic reticulum
0.115
0.120
0.125
0.130
0.135
GeneRatio
Count
26
27
28
29
30
31
p.adjust
1.834130e-23
1.893965e-21
3.769589e-21
5.645212e-21
7.520836e-21
BP- Normal State 5
GeneRatio
BP- Tumor State 5
base-excision repair, AP site formation
positive regulation of receptor binding
pyrimidine deoxyribonucleotide catabolic process
pyrimidine nucleotide catabolic process
pyrimidine deoxyribonucleotide metabolic process
deoxyribonucleotide catabolic process
deoxyribose phosphate catabolic process
thyroid gland development
regulation of receptor binding
cranial nerve morphogenesis
0.200
0.225
0.250
0.275
GeneRatio
Count
1
p.adjust
0.02578749

Supplementary Figure 3: Pathway analysis (GO ontology) of gene-expressions specific to each cell-state in trajectory analysis stratified by malignant and non-malignant sample of origin.

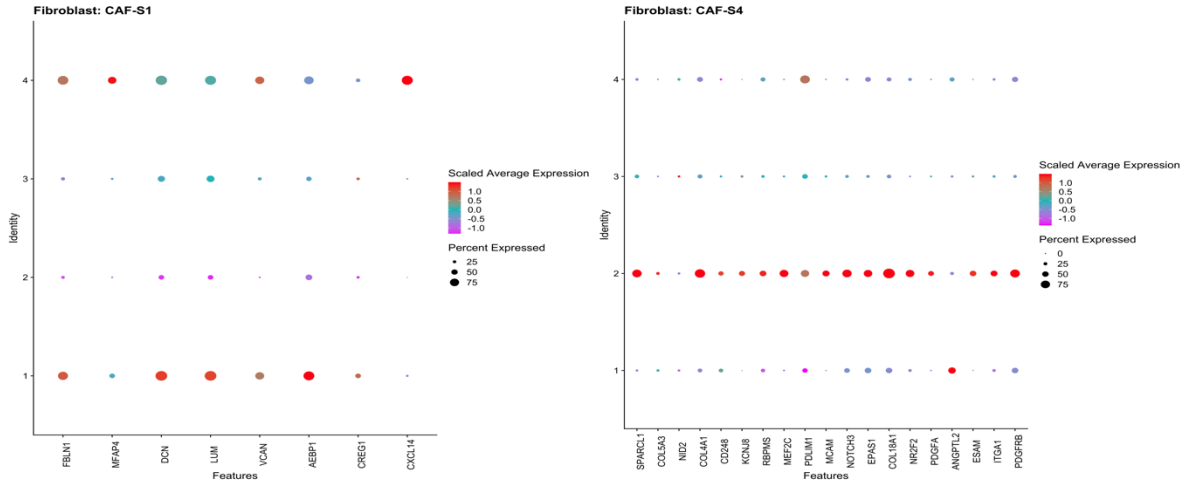


Supplementary Figure 4: Epithelial clustering and differentiation trajectories from a validating cohort (Lee et al. data). a, UMAP clustering of epithelial cells colored by cluster, sample of origin tumor vs. normal tissue status, and microsatellite instability Status (MSI) status. **b,** Differentiation trajectories of epithelial cells colored by differentiation state. **c,** Differentiation trajectory of epithelial cell colored by state and Consensus Molecular Subtypes (CMS) status in *single-cell and in* ***d,*** Differentiation trajectory of epithelial cell colored by state and Consensus Molecular Subtypes (CMS) status in *bulk RNA-seq* using Monocle2 confirming stochastic behavior of tumor epithelial cells (see methods).

a


Clusters
1
2
3
4
Condition
Normal
Tumor
Samples
SMC01-N
SMC01-T
SMC02-T
SMC03-T
SMC04-N
SMC04-T
SMC05-N
SMC05-T
SMC06-N
SMC06-T
SMC07-T
SMC08-N
SMC08-T
SMC09-T
SMC10-N
SMC10-T
SMC11-T
SMC14-T
SMC15-T
SMC16-T
SMC17-T
SMC19-T
SMC20-T
SMC22-T
SMC23-T
SMC24-T
SMC25-T
MSI Status
MSI-H
MSS

b

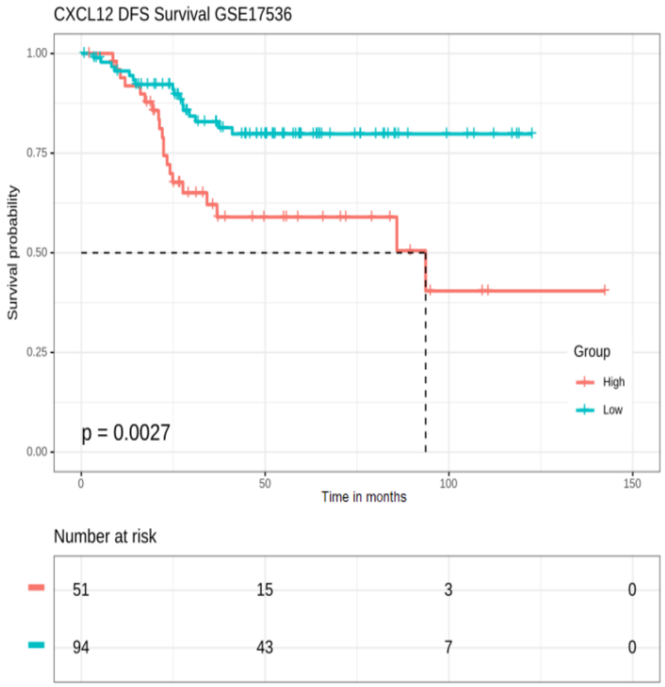

Fibroblast: CAF-S1
Identity
Features
FBLN1
MFAP4
DCN
LUM
VCAN
AEBP1
CREG1
CXCL14
Scaled Average Expression
1.0
0.5
0.0
-0.5
-1.0
Percent Expressed
25
50
75
Fibroblast: CAF-S4
SPARCL1
COL5A3
NID2
COL4A1
CD248
KCNJ8
RBPMS
MEF2C
PDLIM1
MCAM
NOTCH3
EPAS1
COL18A1
NR2F2
PDGFA
ANGPTL2
ESAM
ITGA1
PDGFRB
0

Supplementary Figure 5: Characterization of fibroblasts and their transcriptomic expression patterns (Lee et al. data). **a,** Fibroblasts colored by distinct groups, tumor vs. normal sample, and sample specimen. **b,** Dot plot demonstrating variable expression patterns of subtypes of CAF-S1 and CAF-S4 confirming their relevance in colorectal cancer.

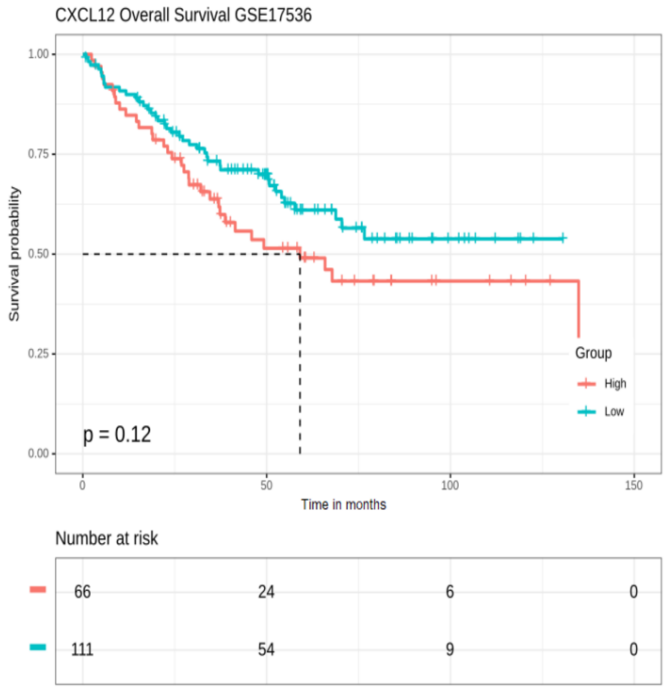
a

GSE17536

DFS



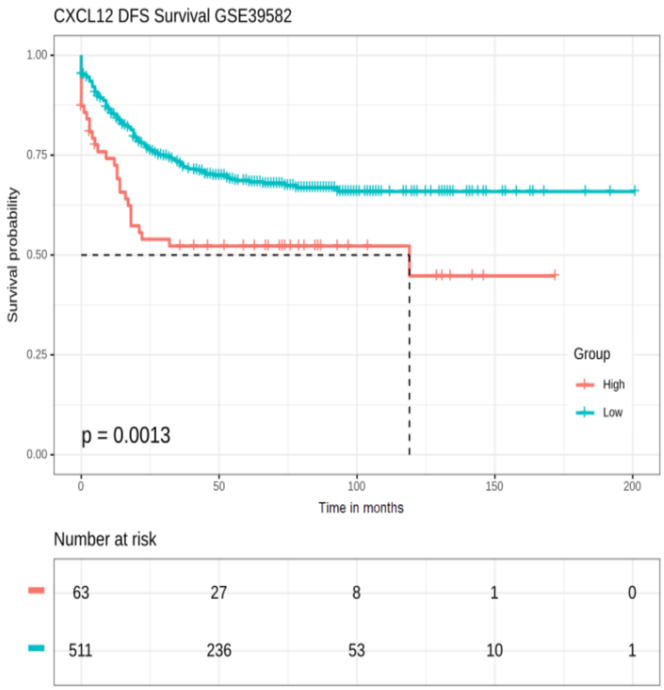
OS



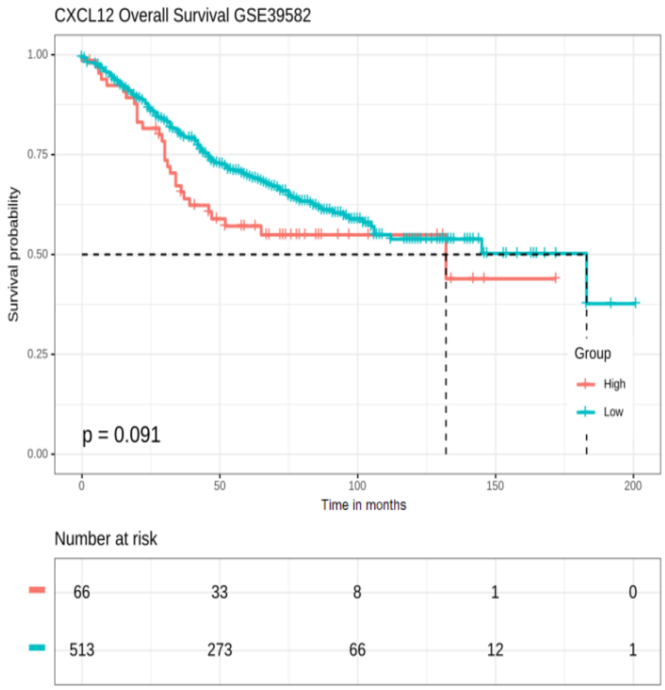
b

GSE39582

DFS



OS



Supplementary Figure 6: a, Kaplan-Meyer analysis of patients with high and low expression levels of CXCL12 (derived from CAF-S1) in the bulk transcriptomic data from GSE17536 (n=177). **b,** Kaplan-Meyer analysis of DFS and OS between patients with high and low expression levels of CXCL12 in the bulk transcriptomic data from GSE39582 (n=585).

a

GSE17536

Cell type	HR	Pvalue
CAF.S1	1.9	0.0044
CAF.S4	2.7	0.0074
CD1C.DENDRITIC.CELLS	0.2	0.045
CD4.Conventional	1.8	0.15
CD4.TH17	0.91	0.71
CD4.TREG	0.22	0.13
CD8.Cytotoxicity	0.99	0.99
CD8.MAIT.CELLS	0.78	0.64
CD8.MEMORY	1.1	0.68
CD8.TRM	0.94	0.83
Endothelial	2.5	0.00065
Epithelial	0.17	0.12
GERMINAL.CENTER	0.32	0.12
INNATE.LYMPHOID.CELLS	1.3	0.37
MEMORY.B.CELLS	1.4	0.12
MONOCYTES	1.2	0.45
NAIVE.CELLS	2.1	0.0029
NK.CELLS	0.94	0.78
PLASMA.CELLS	0.7	0.16
TUMOR.ASSOSIATED.MACROPHAGES	1.4	0.15

b

GSE39582

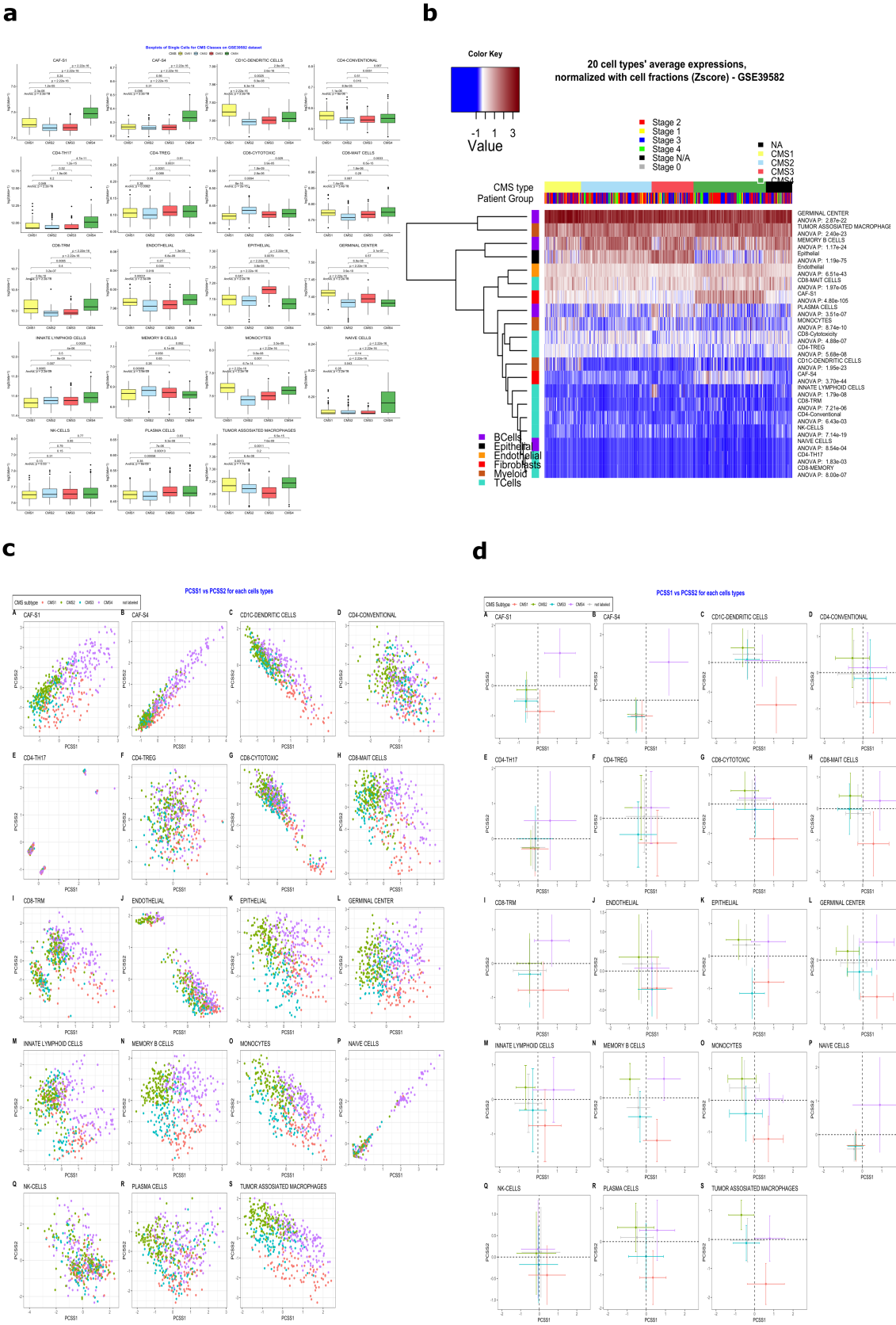
Cell type	HR	Pvalue
CAF.S1	1.4	1.6e−06
CAF.S4	1.6	5.5e−06
CD1C.DENDRITIC.CELLS	1.3	0.13
CD4.Conventional	1.6	2.2e−05
CD4.TH17	0.71	0.28
CD4.TREG	0.75	0.27
CD8.Cytotoxicity	1	0.87
CD8.MAIT.CELLS	1.3	0.07
CD8.MEMORY	1.2	0.039
CD8.TRM	1	0.81
Endothelial	1.5	7.3e−07
Epithelial	0.84	0.56
GERMINAL.CENTER	1.4	0.00013
INNATE.LYMPHOID.CELLS	1.3	0.00085
MEMORY.B.CELLS	1.2	0.015
MONOCYTES	1.2	0.062
NAIVE.CELLS	1.5	8.5e−05
NK.CELLS	1.2	0.023
PLASMA.CELLS	1.1	0.38
TUMOR.ASSOSIATED.MACROPHAGES	1.3	0.00036

c

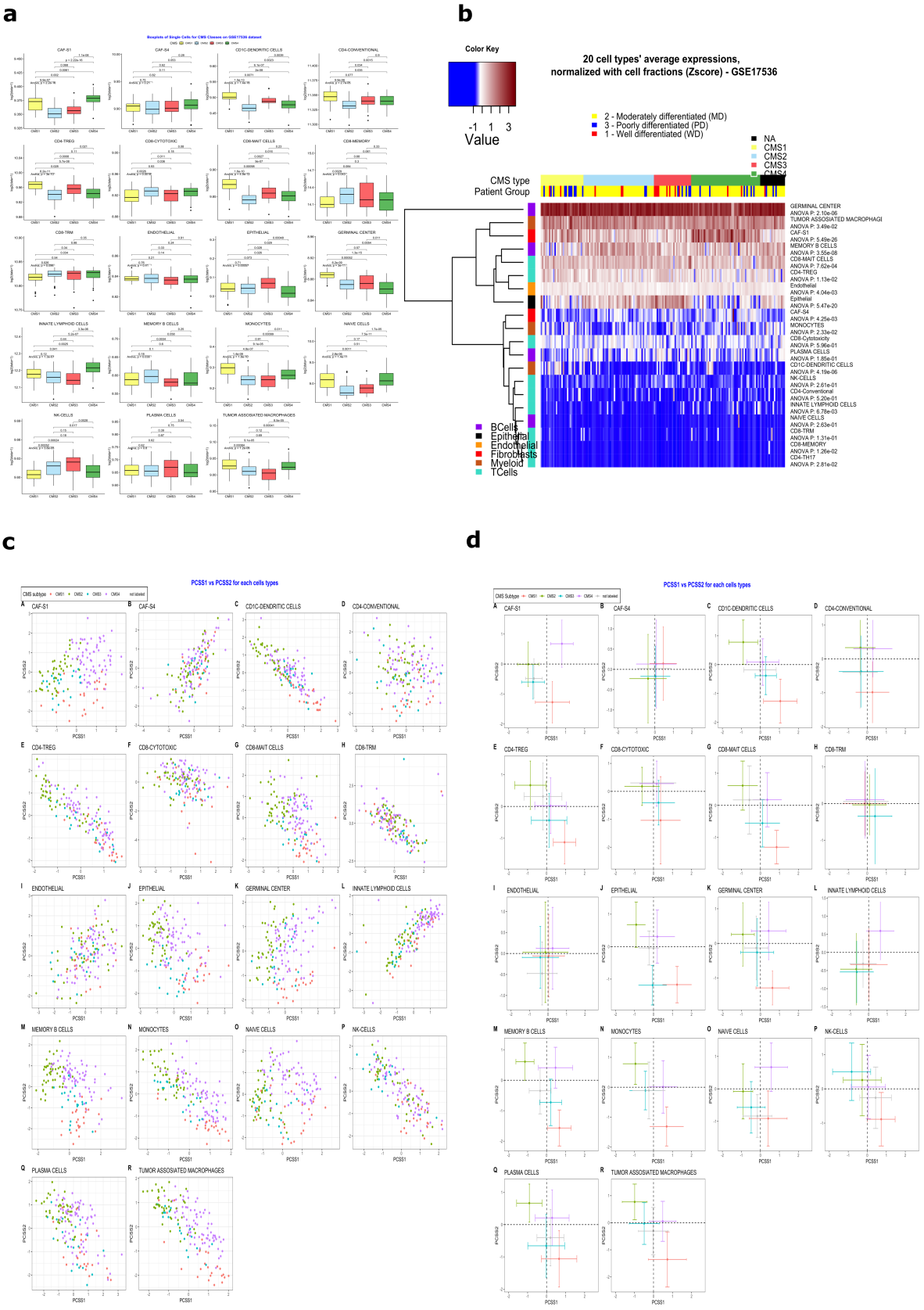
GSE33113

Cell type	HR	Pvalue
CAF.S1	1.7	0.022
CAF.S4	2.1	0.093
CD1C.DENDRITIC.CELLS	2	0.12
CD4.Conventional	1.1	0.57
CD4.TH17	2.5	0.58
CD4.TREG	2.2	0.072
CD8.Cytotoxicity	1.6	0.17
CD8.MAIT.CELLS	1.6	0.26
CD8.MEMORY	2.4	0.0028
CD8.TRM	1.3	0.28
Endothelial	2.2	0.039
Epithelial	1.4	0.79
GERMINAL.CENTER	3.2	0.00027
INNATE.LYMPHOID.CELLS	2.6	0.0059
MEMORY.B.CELLS	1.8	0.05
MONOCYTES	1.7	0.0011
NAIVE.CELLS	3.3	0.0015
NK.CELLS	1.2	0.4
PLASMA.CELLS	1	0.96
TUMOR.ASSOSIATED.MACROPHAGES	1.3	0.27

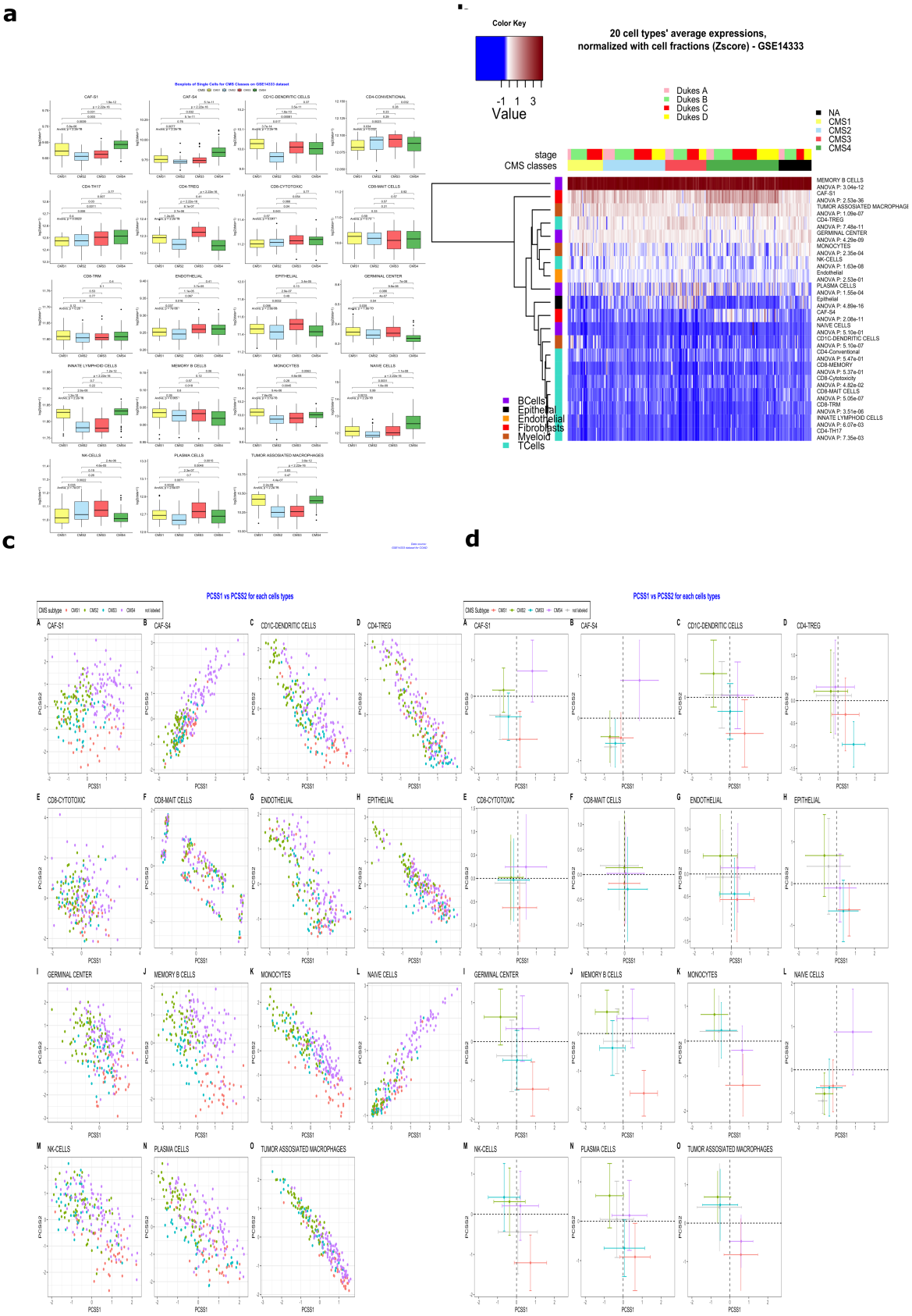
Supplementary Figure 7: Association between relative cell abundance and patient survival from microarray-based datasets. a, GSE17536 (n=177). **b,** GSE39582 (n=585). **c,** GSE33113 (n=96). Note that CAF-S4 is not significant in GSE33113 (p=0.093).


a
Boxplots of Single Cells for CMS Classes on GSE39582 dataset
CMS1 CMS2 CMS3 CMS4
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TH17
CD4-TREG
CD8-CYTOTOXIC
CD8-MAIT CELLS
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
b
Color Key
-1 1 3
Value
20 cell types' average expressions,
normalized with cell fractions (Zscore) - GSE39582
Stage 2
Stage 1
Stage 3
Stage 4
Stage N/A
Stage 0
NA
CMS1
CMS2
CMS3
CMS4
CMS type
Patient Group
GERMINAL CENTER
ANOVA P: 2.87e-22
TUMOR ASSOSIATED MACROPHAGE
ANOVA P: 2.40e-23
MEMORY B CELLS
ANOVA P: 1.17e-24
Epithelial
ANOVA P: 1.19e-75
Endothelial
ANOVA P: 6.51e-43
CD8-MAIT CELLS
ANOVA P: 1.97e-05
CAF-S1
ANOVA P: 4.80e-105
PLASMA CELLS
ANOVA P: 3.51e-07
MONOCYTES
ANOVA P: 8.74e-10
CD8-Cytotoxicity
ANOVA P: 4.88e-07
CD4-TREG
ANOVA P: 5.68e-08
CD1C-DENDRITIC CELLS
ANOVA P: 1.95e-23
CAF-S4
ANOVA P: 3.70e-44
INNATE LYMPHOID CELLS
ANOVA P: 1.79e-08
CD8-TRM
ANOVA P: 7.21e-06
CD4-Conventional
ANOVA P: 6.43e-03
NK-CELLS
ANOVA P: 7.14e-19
NAIVE CELLS
ANOVA P: 8.54e-04
CD4-TH17
ANOVA P: 1.83e-03
CD8-MEMORY
ANOVA P: 8.00e-07
BCells
Epithelial
Endothelial
Fibroblasts
Myeloid
TCells
c
PCSS1 vs PCSS2 for each cells types
CMS subtype
CMS1
CMS2
CMS3
CMS4
not labeled
PCSS1
PCSS2
d
PCSS1 vs PCSS2 for each cells types
CMS Subtype
CMS1
CMS2
CMS3
CMS4
not labeled
PCSS1
PCSS2

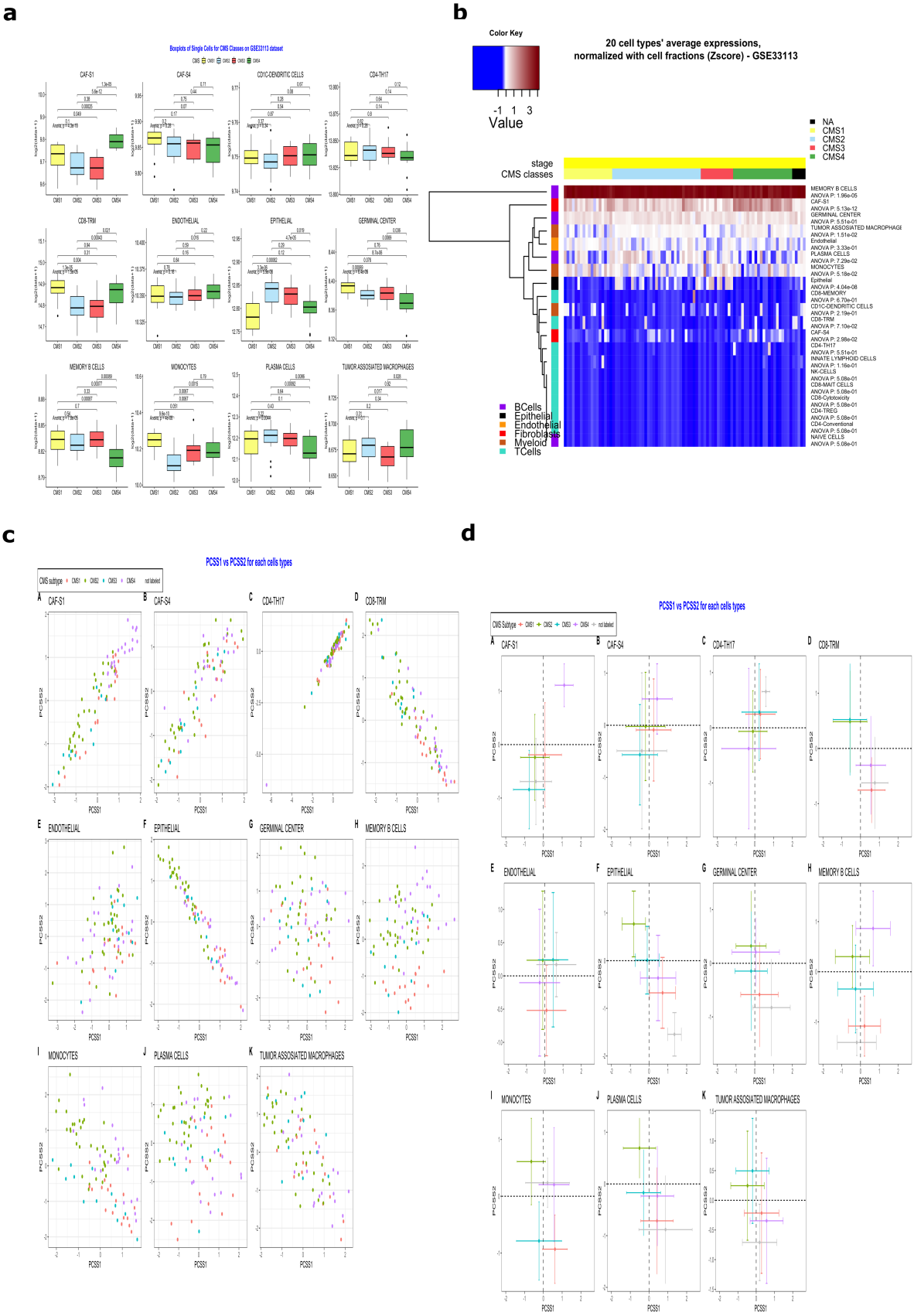
Supplementary Figure 8: Cell abundance prediction for each sample and projected on bulk CMS on GSE39582 (n= 585). a, Boxplots show the distribution of cell types within tumors with varying Consensus Molecular Subtypes(CMS) status. The whiskers depict the 1.5 x IQR. The p-values for one-way ANOVA are shown in the figure. **b**, Deconvolution heatmap of cell type by average expression using CIBERSORTx demonstrating cell type distribution within each CMS status of a single dataset. **c**, 19 cell types show no to little separation reported by CMS. d, All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Markers are colored by the bulk CMS status. Note that, largely, the cell types form a continuum along CMS status and are not clustered in discrete quadrants from one another. Cells are colored by bulk CMS status of sample of origin.


a
Boxplots of Single Cells for CMS Classes on GSE17536 dataset
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TREG
CD8-CYTOTOXIC
CD8-MAIT CELLS
CD8-MEMORY
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
b
Color Key
-1 1 3
Value
20 cell types' average expressions,
normalized with cell fractions (Zscore) - GSE17536
2 - Moderately differentiated (MD)
3 - Poorly differentiated (PD)
1 - Well differentiated (WD)
NA
CMS1
CMS2
CMS3
CMS4
CMS type
Patient Group
GERMINAL CENTER
ANOVA P: 2.19e-06
TUMOR ASSOSIATED MACROPHAGI
ANOVA P: 3.49e-02
CAF-S1
ANOVA P: 5.49e-26
MEMORY B CELLS
ANOVA P: 3.55e-08
CD8-MAIT CELLS
ANOVA P: 7.62e-04
CD4-TREG
ANOVA P: 1.13e-02
Endothelial
ANOVA P: 4.04e-03
Epithelial
ANOVA P: 5.47e-20
CAF-S4
ANOVA P: 4.25e-03
MONOCYTES
ANOVA P: 2.33e-02
CD8-Cytotoxicity
ANOVA P: 5.98e-01
PLASMA CELLS
ANOVA P: 1.85e-01
CD1C-DENDRITIC CELLS
ANOVA P: 4.19e-06
NK-CELLS
ANOVA P: 2.61e-01
CD4-Conventional
ANOVA P: 5.20e-01
INNATE LYMPHOID CELLS
ANOVA P: 6.78e-03
NAIVE CELLS
ANOVA P: 2.63e-01
CD8-TRM
ANOVA P: 1.31e-01
CD8-MEMORY
ANOVA P: 1.26e-02
CD4-TH17
ANOVA P: 2.81e-02
BCells
Epithelial
Endothelial
Fibroblasts
Myeloid
TCells
c
PCSS1 vs PCSS2 for each cells types
CMS subtype
CMS1
CMS2
CMS3
CMS4
not labeled
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TREG
CD8-CYTOTOXIC
CD8-MAIT CELLS
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
PCSS1
PCSS2
d
PCSS1 vs PCSS2 for each cells types
CMS Subtype
CMS1
CMS2
CMS3
CMS4
not labeled
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TREG
CD8-CYTOTOXIC
CD8-MAIT CELLS
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
PCSS1
PCSS2

Supplementary Figure 9: Cell abundance prediction for each sample and projected on bulk CMS on GSE17536 (n= 177). a, Boxplots show the distribution of cell types within tumors with varying Consensus Molecular Subtyping (CMS) status. The whiskers depict the 1.5 x IQR. The p-values for one-way ANOVA are shown in the figure. **b**, Deconvolution heatmap of cell type by average expression using CIBERSORTx demonstrating cell type distribution within each CMS status of a single dataset. **c**, 18 cell types show no to little separation reported by CMS. **d**, All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Markers are colored by the bulk CMS status. Note that, largely, the cell types form a continuum along CMS status and are not clustered in discrete quadrants from one another. Cells are colored by bulk CMS status of sample of origin.

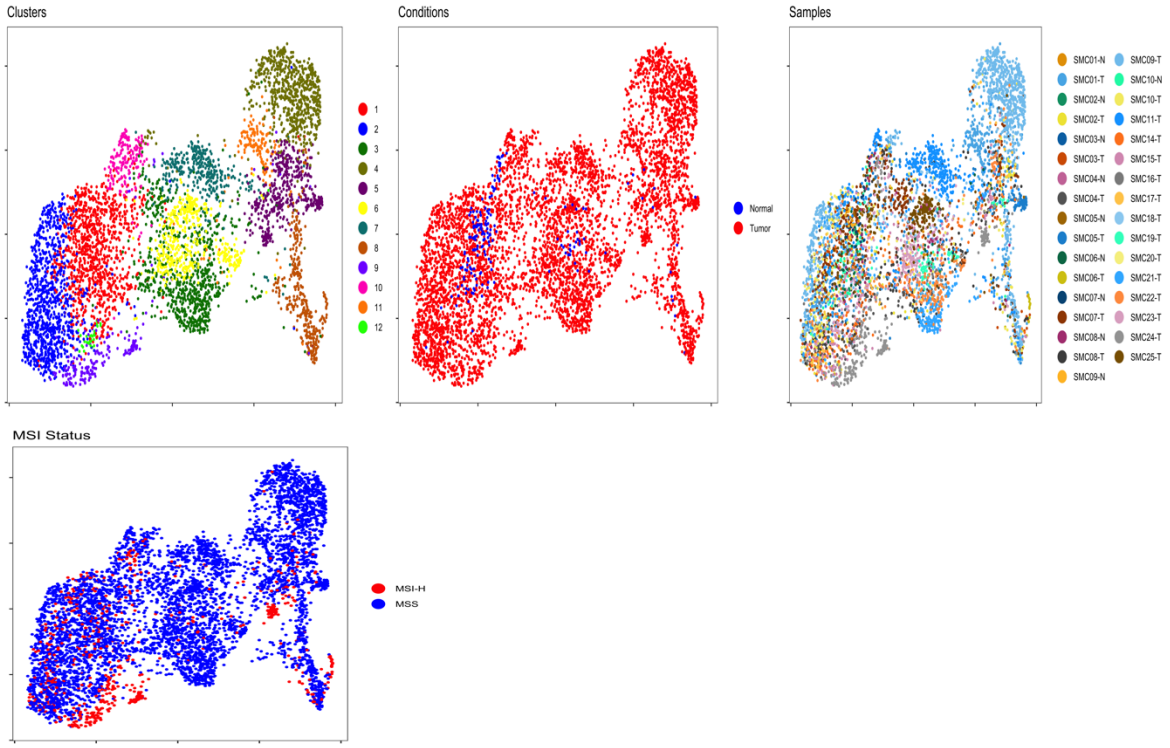

a
Boxplots of Single Cells for CMS Classes on GSE14333 dataset
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-CONVENTIONAL
CD4-TH17
CD4-TREG
CD8-CYTOTOXIC
CD8-MAIT CELLS
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
INNATE LYMPHOID CELLS
MEMORY B CELLS
MONOCYTES
NAIVE CELLS
NK-CELLS
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
CMS1
CMS2
CMS3
CMS4
Color Key
-1 1 3
Value
20 cell types' average expressions,
normalized with cell fractions (Zscore) - GSE14333
Dukes A
Dukes B
Dukes C
Dukes D
NA
CMS1
CMS2
CMS3
CMS4
stage
CMS classes
MEMORY B CELLS
ANOVA P: 3.04e-12
CAF-S1
ANOVA P: 2.53e-36
TUMOR ASSOSIATED MACROPHAGI
ANOVA P: 1.09e-07
CD4-TREG
ANOVA P: 7.48e-11
GERMINAL CENTER
ANOVA P: 4.29e-09
MONOCYTES
ANOVA P: 2.35e-04
NK-CELLS
ANOVA P: 1.63e-08
Endothelial
ANOVA P: 2.53e-01
PLASMA CELLS
ANOVA P: 1.55e-04
Epithelial
ANOVA P: 4.89e-16
CAF-S4
ANOVA P: 2.08e-11
NAIVE CELLS
ANOVA P: 5.10e-01
CD1C-DENDRITIC CELLS
ANOVA P: 5.10e-07
CD4-Conventional
ANOVA P: 5.47e-01
CD8-MEMORY
ANOVA P: 5.37e-01
CD8-Cytotoxicity
ANOVA P: 4.82e-02
CD8-MAIT CELLS
ANOVA P: 5.05e-07
CD8-TRM
ANOVA P: 3.51e-06
INNATE LYMPHOID CELLS
ANOVA P: 6.07e-03
CD4-TH17
ANOVA P: 7.35e-03
BCells
Epithelial
Endothelial
Fibroblasts
Myeloid
TCells
c
d
PCSS1 vs PCSS2 for each cells types
CMS subtype
CMS1
CMS2
CMS3
CMS4
not labeled
PCSS1
PCSS2

Supplementary Figure 10: Cell abundance prediction for each sample and projected on bulk CMS on GSE14333 (n=290). **a**, Boxplots show the distribution of cell types within tumors with varying Consensus Molecular Subtypes (CMS) status. The whiskers depict the 1.5 x IQR. The p-values for one-way ANOVA are shown in the figure. **b**, Deconvolution heatmap of cell type by average expression using CIBERSORTx demonstrating cell type distribution within each CMS status of a single dataset **c**, 19 cell types show no to little separation reported by CMS. **d**, All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Markers are colored by the bulk CMS status. Note that, largely, the cell types form a continuum along CMS status and are not clustered in discrete quadrants from one another. Cells are colored by bulk CMS status of sample of origin.

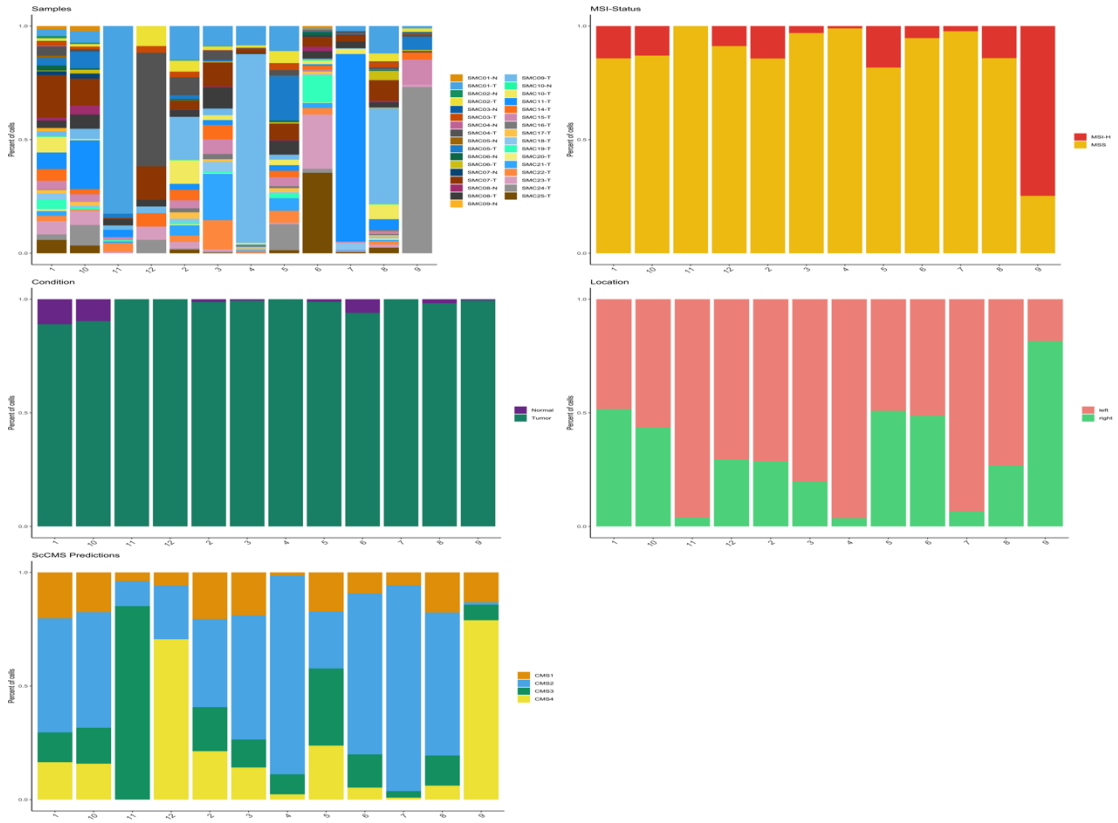

a
Boxplots of Single Cells for CMS Classes on GSE33113 dataset
CAF-S1
CAF-S4
CD1C-DENDRITIC CELLS
CD4-TH17
CD8-TRM
ENDOTHELIAL
EPITHELIAL
GERMINAL CENTER
MEMORY B CELLS
MONOCYTES
PLASMA CELLS
TUMOR ASSOSIATED MACROPHAGES
b
Color Key
-1 1 3
Value
20 cell types' average expressions,
normalized with cell fractions (Zscore) - GSE33113
NA
CMS1
CMS2
CMS3
CMS4
stage
CMS classes
BCells
Epithelial
Endothelial
Fibroblasts
Myeloid
TCells
c
PCSS1 vs PCSS2 for each cells types
CMS subtype
CMS1
CMS2
CMS3
CMS4
not labeled
PCSS1
PCSS2
d
PCSS1 vs PCSS2 for each cells types
CMS Subtype
CMS1
CMS2
CMS3
CMS4
not labeled
PCSS1
PCSS2

Supplementary Figure 11: Cell abundance prediction for each sample and projected on bulk CMS on GSE33113 (n=96). a, Boxplots show the distribution of cell types within tumors with varying Consensus Molecular Subtypes (CMS) status. The whiskers depict the 1.5 x IQR. The p-values for one-way ANOVA are shown in the figure. **b**, Deconvolution heatmap of cell type by average expression using CIBERSORTx demonstrating cell type distribution within each CMS status of a single dataset **c**, 19 cell types show no to little separation reported by CMS. **d**, All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Markers are colored by the bulk CMS status. Note that, largely, the cell types form a continuum along CMS status and are not clustered in discrete quadrants from one another. Cells are colored by bulk CMS status of sample of origin.

a

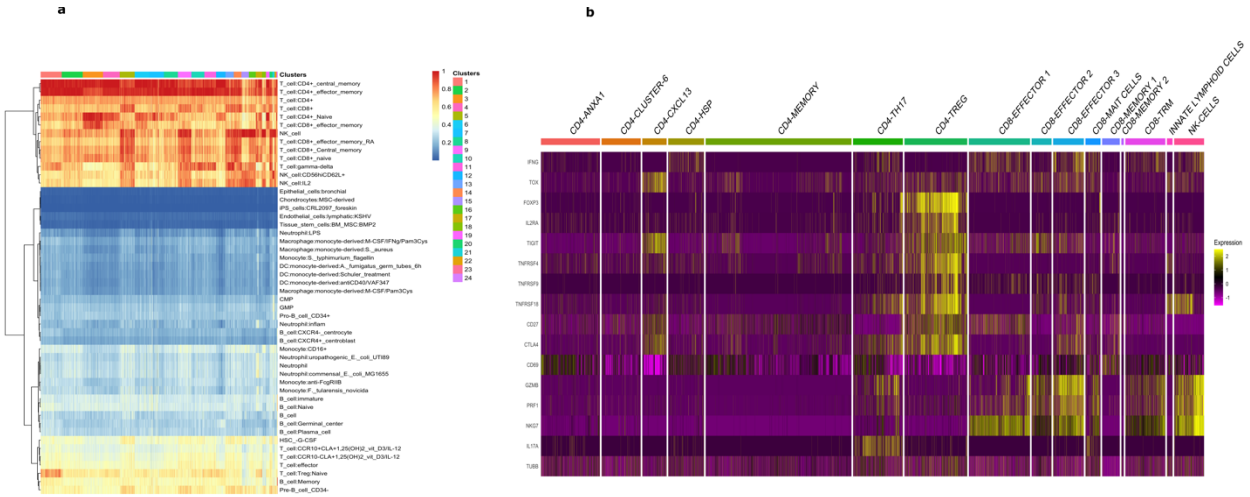

Clusters
1
2
3
4
5
6
7
8
9
10
11
12
Conditions
Normal
Tumor
Samples
SMC01-N
SMC01-T
SMC02-N
SMC02-T
SMC03-N
SMC03-T
SMC04-N
SMC04-T
SMC05-N
SMC05-T
SMC06-N
SMC06-T
SMC07-N
SMC07-T
SMC08-N
SMC08-T
SMC09-N
SMC09-T
SMC10-N
SMC10-T
SMC11-T
SMC14-T
SMC15-T
SMC16-T
SMC17-T
SMC18-T
SMC19-T
SMC20-T
SMC21-T
SMC22-T
SMC23-T
SMC24-T
SMC25-T
MSI Status
MSI-H
MSS

b


Samples
MSI-Status
MSI-H
MSS
Condition
Normal
Tumor
Location
left
right
ScCMS Predictions
CMS1
CMS2
CMS3
CMS4
Percent of cells

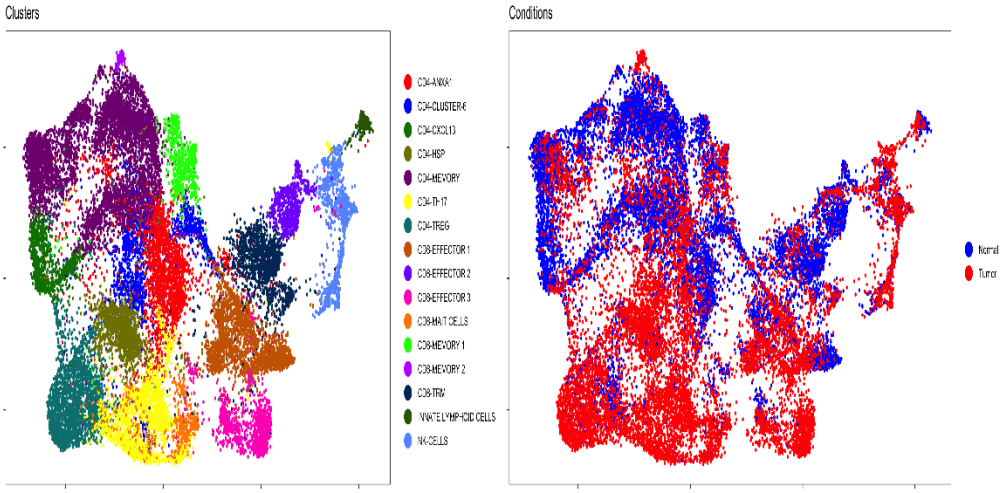
13

Supplementary Figure 12: T/NK cell data from Lee et al. a, Reclustering of T/NK cells and coloring by clusters, tumor vs. normal status, MSI status, and samples. **b**, Bar chart representation of cells colorized by our samples of origin, MSI status, tumor status, colonic location, Consensus Molecular Subtypes (CMS) status.

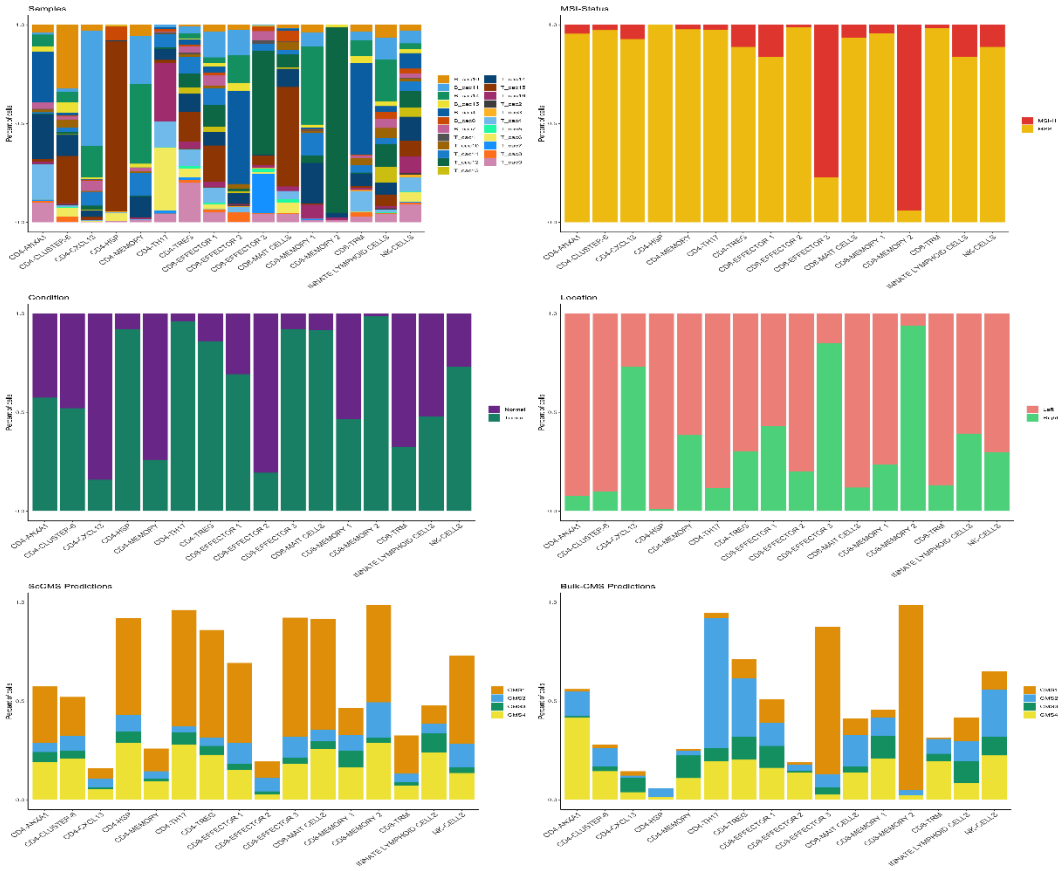


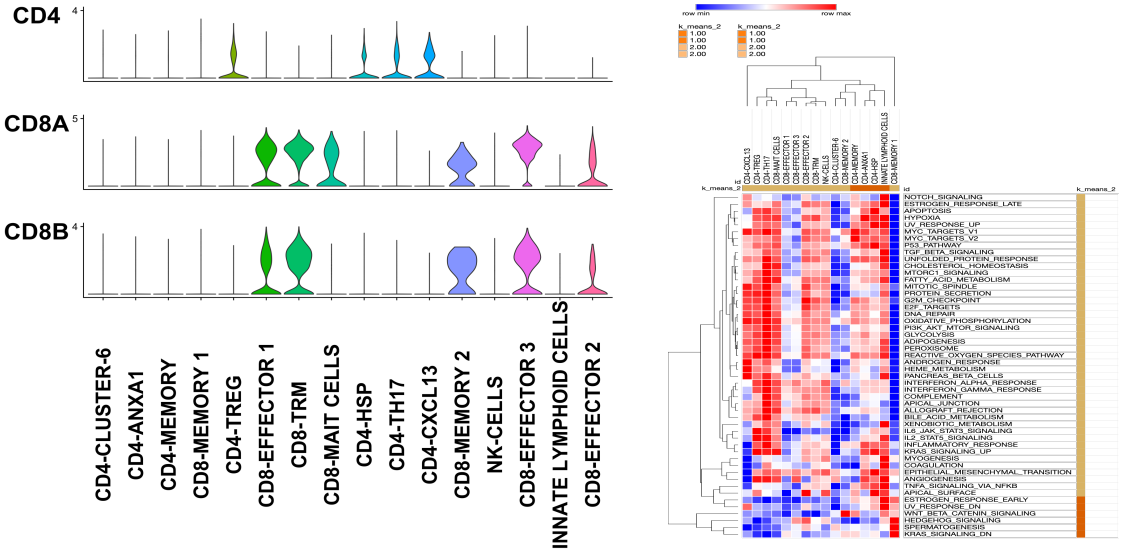
Supplementary Figure 13: T cell expression analyses across T/NK cells. a, SingleR heatmap cell type identification within each cluster. Note: doublets were removed from the further analysis. **b,** T/NK cell gene specific expression to identify T cell heterogeneity using published literature (see methods).

a


Clusters
Conditions
CD4-NKG4
CD4-CLUSTER-6
CD4-CXCL13
CD4-HSP
CD4-MEMORY
CD4-TH-17
CD4-TREG
CD8-EFFECTOR 1
CD8-EFFECTOR 2
CD8-EFFECTOR 3
CD8-MAIT CELLS
CD8-MEMORY 1
CD8-MEMORY 2
CD8-TRM
NAIVE LYMPHOID CELLS
NK-CELLS
Normal
Tumor

b

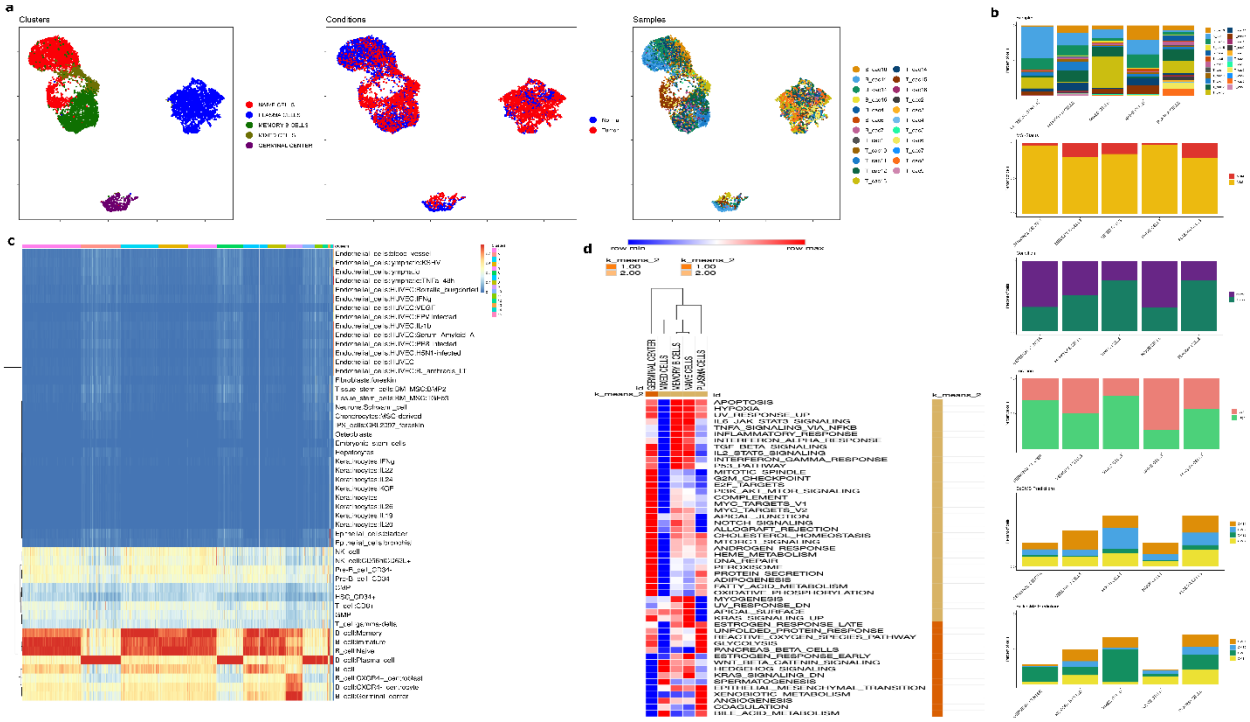

Samples
MSI-Status
Condition
Location
ScCMS Predictions
Bulk-CMS Predictions
Percent of cell
Normal
Tumor
CMS1
CMS2
CMS3
CMS4



Extended Data Figure 1. T cell identification and characterization. a, UMAP reclustering of T cells colored by cell phenotype, tissue malignancy status, and sample of origin. **b,** Bar plot depiction of the proportion of cells colored by sample of origin, MSI status, malignancy status, tumor location, scCMS status, and bulk CMS status. **c,** Violin plots showing the differential expression of T cell-specific marker genes between CD4 and CD8 phenotypes. **d,** Heatmap of the Hallmark pathway analysis for the T cell compartment. scCMS, single-cell consensus molecular subtyping; bulk CMS, consensus molecular subtyping on Bulk RNA-seq data.

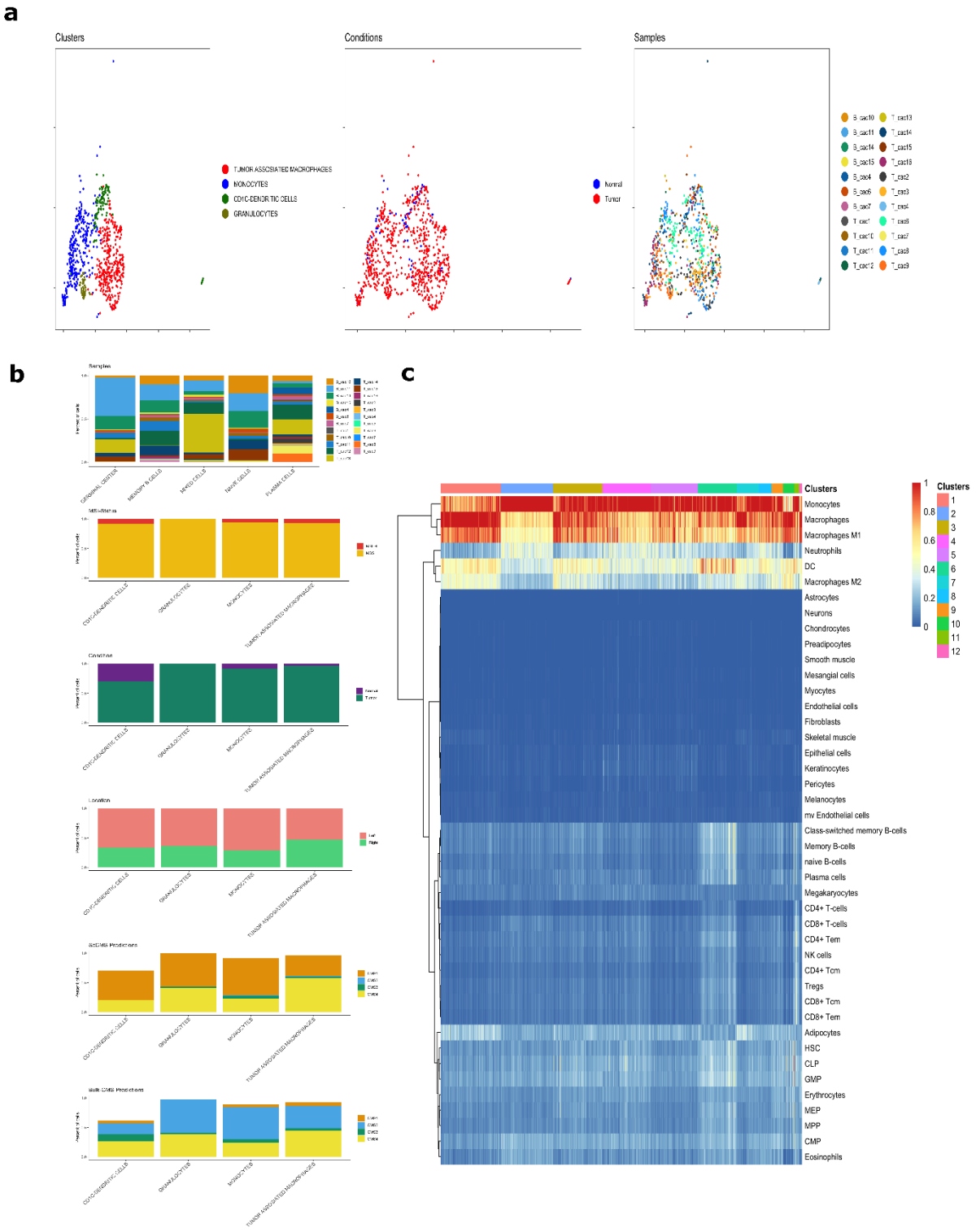
Description: We reclustered and analyzed 22,525 cells from both tumors and adjacent, normal tissue samples and identified 11 CD4+ T cell and 12 CD8+ T cell clusters. We used known canonical markers and published expression signatures to identify T cell states for further analysis (see methods). We identified conventional CD4+ T cells, CD4+ Tregs, CD8+ (naive/memory, cytotoxic, resident memory, and MAIT cells), NK cells, and innate lymphoid cells (ILC). Among the conventional CD4+ T cells, we identified the central memory/naive like-state (CCR7+, SELL+, and TCF7+) enriched in non-malignant samples. In contrast, Th17 cells expressing IL-17, known as critical anti-tumor effectors, were enriched in tumor samples. CD4+ Tregs (FOXP3+) expressing immune checkpoint markers and costimulatory molecules were among the most abundant T cells in the colorectal TME compared to non-malignant tissue. Among the CD8+ T cell states, CD8+ cytotoxic cells were distributed across three clusters that we labeled CD8+ effector 1, CD8+ effector 2, and CD8+ effector 3. CD8+ effector clusters expressed cytotoxicity genes and chemokines as previously described in other tumor types. CD8 effector3 was predominantly enriched in MSI-H CRC patients and represented 77% of the total CD8+ effector 3 population among the 2 MSI-H CRC samples. This cluster expressed ITGAE, LAYN, CXCL13, and T cell exhaustion markers (LAG3, HAVCR2, and CD96), possibly explaining this CD8+ cell state's role in the response to immune checkpoint inhibitors in MSI-H colorectal tumors. Gene-set enrichment of CD8+ effectors further confirmed their distinct states. CD8+ effector 2 was a proliferative cluster with MYC activity, NOTCH activation, and EF2 targets. CD8+ effector(s), CD8+ MAIT cells, and NK cells were enriched in tumors, whereas Tissue-resident memory (Trm) cells were depleted in tumor tissue. Trm induction was recently

seen to enhance cancer vaccine efficacy in other tumors, suggesting a possible therapeutic target in CRC.



Extended Data Figure 2. **Reclustering of B cells with characterization of clusters and phenotypes. a,** UMAP depiction of B cell reclustering colored by cluster, malignancy status, and sample of origin. **b,** Bar plot depiction of the proportion of cells within each B cell phenotype colored by sample of origin, MSI status, malignancy status, tumor location, scCMS and bulk CMS. **c,** SingleR heatmap demonstration of B cell distribution within each cluster. **d,** B cell Hallmark pathway analysis by phenotype. scCMS, single-cell consensus molecular subtyping; bulk CMS, consensus molecular subtyping on Bulk RNA-seq data.

Description: To illustrate characteristics of B cells in CRC we reclustered 9,289 B cells that clearly identified naive cells, memory cells, plasma cells, and germinal center (GC) B cells. All B cell subtypes from the CRC TME and the non-malignant colonic tissue clustered together exhibiting transcriptional similarity among non-tumor and tumor-derived cells. Memory B cells and plasma cells were enriched in tumors, while naive and GC B cells were enriched in non-malignant tissue.


a
Clusters
TUMOR ASSOCIATED MACROPHAGES
MONOCYTES
CD1C-DENDRITIC CELLS
GRANULOCYTES
Conditions
Normal
Tumor
Samples
b
Samples
Conditions
Location
c
Clusters
Monocytes
Macrophages
Macrophages M1
Neutrophils
DC
Macrophages M2
Astrocytes
Neurons
Chondrocytes
Preadipocytes
Smooth muscle
Mesangial cells
Myocytes
Endothelial cells
Fibroblasts
Skeletal muscle
Epithelial cells
Keratinocytes
Pericytes
Melanocytes
mv Endothelial cells
Class-switched memory B-cells
Memory B-cells
naive B-cells
Plasma cells
Megakaryocytes
CD4+ T-cells
CD8+ T-cells
CD4+ Tem
NK cells
CD4+ Tcm
Tregs
CD8+ Tcm
CD8+ Tem
Adipocytes
HSC
CLP
GMP
Erythrocytes
MEP
MPP
CMP
Eosinophils

127 **Extended Data Figure 3**. **Reclustering of the myeloid cell compartment. a,** UMAP depiction
128 of myeloid cell reclustering colored by subtype, malignancy status, and sample of origin. **b,** Bar
129 plot depiction of proportion of cells within each myeloid cell phenotype colored by sample of
130 origin, MSI status, malignancy status, tumor location, scCMS and bulk CMS. **c,** SingleR
131 heatmap demonstration of myeloid cell distribution within each cluster. scCMS, single-cell
132 consensus molecular subtyping; bulk CMS, consensus molecular subtyping on Bulk RNA-seq
133 data.

134 **Description:** We reclustered 819 myeloid cells and identified CD1C+ dendritic cells, tumor-
135 associated macrophages (TAM and MRC1+), monocytes (S100A8+), and granulocyte clusters.
136 We recovered key cell types including M2 polarized macrophages, as seen in other tumor types.
137 Monocytes revealed proinflammatory phenotypes (1L1B, S100A8, and S100A9), while TAM
138 showed anti-inflammatory signatures (APOE, SEPP1, and CD163) consistent with the role of
139 TAM in immune suppression and cancer progression (see methods).

140
141
142