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

Ateeq M. Khaliq[1,*], Zeyneb Kurt[2,*], Miles W. Grunvald[1,*], Cihat Erdogan[3], Sevgi S. Turgut[4], Tim Rand[5], Sonal Khare[5], Jeffrey 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[6], Kiran Turaga[6], Milan Radovich[7], Vineet Gupta[1], Ram Al-Sabti[1], Sheeno Thyparambil[8], Mia A. Levy[1], Janakiraman Subramanian[9], Timothy M. Kuzel[1], Anguraj Sadanandam[10], Arif Hussain[12], Bassel El-Rayes[12], Ameen A. Salahudeen[5✉], 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]Yildiz Technical University, Istanbul, Turkey.

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

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

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

[8]mProbe Inc. Rockville, Maryland, USA.

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

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

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

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

INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and a leading cause of cancer-related mortality worldwide[1,2]. Approximately one-third of patients experience disease relapse following curative-intent surgical resection and chemotherapy[3]. Despite the high incidence and mortality of advanced CRC, few effective therapies have been approved in the past several decades[4]. One barrier to the development of efficacious therapeutics is the biological heterogeneity of CRC and its variable clinical course. While landmark studies from The Cancer Genome Atlas (TCGA) have defined the somatic mutational landscape within CRC, several studies have shown that stromal signatures, including fibroblasts and cytotoxic T cells, are likely the main drivers of clinical outcomes[5–9]. These findings suggest that the clinical phenotypes of CRC and by extension, its tumor biology is shaped by a complex niche of heterotypic cell interactions within the tumor microenvironment (TME).

Bulk gene expression analyses by several independent groups have identified distinct CRC subtypes[10–12]. Based on bulk transcriptomic signatures, an international consortium published the Consensus Molecular Subtypes (CMS) that classified CRC into CMS1 (MSI immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal) subtypes[12]. Unfortunately, associations between CMS and meaningful therapeutic responses to specific agents have been inconsistent across studies[13–16]. Further, there has been a lack of concordance between primary and metastatic CRC tumors within the CMS framework, limiting its overall utility in clinical decision making[16–18]. Thus, an improved CMS classification, or an alternative classification system, is needed to improve clinical utility.

To overcome the limitations of bulk-RNA sequence profiling, we utilized single-cell RNA sequencing (scRNA-seq) to more thoroughly evaluate the CRC subtypes at the molecular level, including within the context of the currently defined CMS classification. We dissected heterotopic cell states of tumor epithelia and stromal cells, including a cancer-associated fibroblast (CAF) population. The CAF population's clinical and prognostic significance became apparent when CAF signatures were applied to large, independent CRC transcriptomic cohorts.

RESULTS

We profiled sixteen primary CRC tumor tissue samples and eight adjacent, normal, colonic tissue samples (24 in total) using droplet-based, scRNA-seq. Altogether, we captured and retained 49,589 high-quality single cells after performing quality control for downstream analysis (**Fig. 1a, Supplementary Table 1**). All scRNA-seq data were merged and normalized to identify robust discrete clusters of epithelial cells (*EPCAM+*, *KRT8+*, and *KRT18+*), fibroblasts (*COL1A2+*), endothelial cells (*CD31+*), T cells (*CD3D+)*, B cells *(CD79A+)*, and myeloid cells (*LYZ+*) using canonical marker genes **(Fig. 1b-c).** Additionally, each cell type compartment was analyzed separately. Clustree (v0.4.1) and manual review of differentially expressed genes in each subcluster were studied to choose the best cluster resolution without cluster destabilization (see methods)[19]. Cell population designation was chosen by specific gene expression, and SingleR was also utilized for unbiased cell type recognition (see methods)[20–23].

In addition to cancer cells, we identified diverse TME cell phenotypes, including fibroblasts subsets (*CAF-S1* and *CAF-S4*), endothelial cells, CD4+ subsets (naïve/memory, Th17, and Tregs), CD8+ subsets (naïve/memory, cytotoxic, tissue-resident memory, and Mucosa-Associated Invariant (MAIT) cells), NK cells, innate lymphoid cell (ILC) types, B cell phenotypes (naïve, memory, germinal center, and plasma cells), and monocyte lineage phenotypes (C1DC+ dendritic cells, proinflammatory monocytes [IL1B, IL6, S100A8, and S100A9]), and M2 polarized anti-inflammatory [*CD163*, *SEPP1*, *APOE*, and *MAF]*), tumor-associated macrophages (TAMs) (**Fig. 1d, Extended Data Figs. 1-3, Extended Data Tables 1-4**)[20–22,24].

We also profiled an independent CRC dataset for validation and retained 31,863 cells after quality control and strict filtering of cells expressing hybrid markers[25]. The re-clustering of individual compartments further refined our analysis, and cells that expressed hybrid or distinct lineage markers within a cluster were removed from the downstream analysis (**Supplementary Fig. 4-5, 9-10**) (see methods). Thus, a total of 81,242 high-quality cells were profiled to produce a single-cell map of 39 colorectal cancer patients. The results of the primary CRC cohort (49,859 single-cells) are available at the Colon Cancer Atlas (*www.websiteinprogress.com*).

Malignant colon cancer reveals tumor epithelial cell subclonal heterogeneity and stochastic behavior.

We detected 8,965 tumor and benign epithelial cells (*EPCAM+*, *KRT8+*, and *KRT18+*) and, on re-clustering, produced 17 epithelial clusters (designated C1 to C17) (**Fig. 2a-b**). Tumor cells were confirmed to be of malignant origin by inferring chromosomal copy number alterations (**Supplementary Fig. 1**). Clusters were chiefly influenced by colonic epithelial markers, including those for stemness (*LGR5, ASCL2, OLFM4,* and *STMN1*), enterocytes (*FABP1* and *CA2*), goblet cells (*ZG16, MUC2, SPINK4,* and *TFF3*), and enteroendocrine cells (*PYY* and *CHGA*) (**Supplementary Table 2**). Tumor cells exhibited a high degree of de-differentiated state of plasticity, possibly accounting for lasting cancer growth (**Supplementary Fig. 2b**)[26]. Distinct tumor-derived clusters were predominantly patient-specific, reflecting a high degree of inter-patient tumoral cell heterogeneity **(Fig. 1d).** In contrast, epithelial populations derived from normal colon tissue samples across multiple patients clustered together, a pattern observed in previous studies confirming both normal tissue homeostasis and limited sample batch effects (**Fig. 1d**)[27,28].

We next aimed to identify gene expression programs shared across these clusters using hallmark pathway analysis[29]. A strong overlap was observed for multiple pathways such as activation of inflammatory, epithelial-mesenchymal transformation (EMT), immune response, and metabolic pathways (**Fig. 2b**). Interestingly, high microsatellite instability (MSI-H) and microsatellite stable (MSS) CRC tumors, considered clinically separate entities, demonstrated similar pathway program activation within the tumor epithelial populations **(Fig. 2b).** Some clusters also showed activation of unique pathways such as activation of apical junctions and angiogenesis (C6), hypoxia and fatty acid metabolism (C11) and Notch signaling and DNA repair (C14), among others **(Fig. 2b).** However, MSI-H tumors differed from MSS tumors based on immune cell infiltration (**Extended Data Fig. 1**).

Since intratumoral heterogeneity is recognized as a key mechanism contributing to drug resistance, cancer progression, and recurrence, we next focused on dissecting potential transcriptomic states to identify heterogeneity within each tumor[30–32]. We found that each tumor specimen contained 2-10 distinct tumor epithelial clusters (**Fig. 1d**). Gene set variation analysis (GSVA) was performed on cells from individual tumor samples and illustrated the sub-clonal transcriptomic heterogeneity within each specimen (**Supplementary Fig. 2c**)[33]. Clusters identified in individual pathway analysis demonstrated the up- or down-regulation of crucial metabolic and oncogenic pathways between samples, suggesting wide phenotype variations between cells from the same tumor[34].

Given the evidence of intratumoral epithelial heterogeneity, we next performed trajectory inference using pseudotime analysis to identify potential alignments or lineage relationships (i.e., right versus left-sided CRC), CMS classification, or MSI status[35,36]. This analysis also served as a control for inter-patient genomic heterogeneity and provided an orthogonal strategy to confirm the transcriptomic trends we identified. We detected five molecular states (S1n/t to S5n/t) with malignant and normal epithelial cells intermixed and aligned along a common transcriptional trajectory **(Fig. 2c)**[37,38]. In both normal and tumor cells, each transcriptional state pathway activation was shared and related to colon epithelial function of nutrient absorption, maintaining the colon homeostasis, and the activation of cancer-related pathways such as apoptosis and cell development[39]. However, these cell states showed differential gene enrichment activation between normal and tumor cells **(Supplementary Fig. 3, Supplementary Table 3)**.

Additionally, tumor cells showed upregulation of embryogenesis (S2t), consistent with previous findings that tumor cells revert to their embryological states in cancer development (**Supplementary Table 3**)[40].

Interestingly, there were no significant associations with anatomic location, CMS classification, or MSI status within our dataset or an independent dataset of 31,383 single cells (**Fig. 2d**, **Supplementary Fig. 4**)[25]. Hence, in our analysis, CRC oncogenesis represent the hijacking of the normal epithelial differentiation pathways coupled with the acquisition of embryonic pathways[37,40].

CRC-associated fibroblasts in the tumor microenvironment exhibit diverse phenotypes, and specific subtypes are associated with poor prognosis.

We next focused on CRC TME subpopulations. 819 high-quality fibroblasts were re-clustered into eight clusters, and then phenotypically classified into two major subtypes to assess for further CAF heterogeneity. These phenotypic subtypes were found to be immunomodulatory CAF-S1 (*PDGFRA+* and *PDPN+*) and contractile CAF-S4 (*RGS5+* and *MCAM+*) (**Fig. 3a-b, Supplementary Table 4**)[41,42]. This fibroblast cluster dichotomy was also observed in the independent CRC patient scRNA-seq dataset of 31,383 cells (**Supplementary Fig. 5**)[25].

The CAF-S1 and CAF-S4 subtypes showed striking resemblances to the mCAF (extracellular matrix) and vCAF (vascular) fibroblast subtypes, respectively, as previously described in a mouse breast cancer model[43]. Most clusters were found in multiple patients, albeit in varying proportions, signifying shared patterns in CAF transcriptomic programs between patients. Fibroblasts derived from MSI-H tumors were distributed similarly throughout these clusters (**Fig. 1d**).

CAF-S1 cells exhibited high chemokine expressions such as *CXCL1*, *CXCL2*, *CXCL12*, *CXCL14*, and immunomodulatory molecules including *TNFRSF12A* (**Supplementary Table S4)**. Additionally, CAF-S1 cells expressed extracellular matrix genes including matrix-modifying enzymes (*LOXL1* and *LOX*)[43]. To determine this population's functional significance, we compared the CAF-S1 population transcriptomes to those described recently in breast cancer, lung cancer, and head and neck cancer[44]. We recovered five CAF-S1 subtypes that included ecm-myCAF (extracellular; *GJB2*), IL-iCAF (growth factor, *TN*F and interleukin pathway; *SCARA5*), detox-iCAF (detoxification and inflammation; *ADH1B*), wound-myCAF (collagen fibrils and wound healing; *SEMA3C*), and TGFβ-myCAF (*TGF-β* signaling and matrisome; *CST1*, *TGFb1*), which were previously divided into two major subtypes: iCAF and myCAF (**Fig. 3c**). Among these five subtypes, ecm-myCAF and TGFβ-myCAF are known to correlate with immunosuppressive environments and are enriched in tumors with high regulatory T lymphocytes (Tregs) and depleted CD8+ lymphocytes. Additionally, these subtypes are associated with primary immunotherapy resistance in melanoma and lung cancer[44].

The CAF-S4 population expressed pericyte markers (*RGS5+*, *CSPG4+*, and *PDGFRA+*), *CD248 (endosialin)*, and *EPAS1 (HIF2-α),* that this particular CAF subtype is vessel-associated, with hypoxia potentially contributing to invasion and metastasis, as has been shown in another study **(Fig. 3a-b, Supplementary Table 4)**[43]. CAF-S4 clustered into the immature phenotype (*RGS5+*, *PDGFRB+*, and *CD36+)* and the differentiated myogenic subtype (*TAGLN+* and *MYH11+*) **(Supplementary Table 4)**[42].

Given the correlation between CMS4 and fibroblast infiltration, we next sought to test the existence of CAF-S1 and CAF-S4 signatures in bulk transcriptomic data and their association with clinical outcomes[12]. To this end, we interrogated and carried out a meta-analysis of eight colorectal cancer transcriptomic datasets comprising 1,584 samples and confirmed the presence of CAF-S1 and CAF-S4 gene signatures in CRC and other cancer types (**Fig. 4a**). We detected a strong and positive correlation between specific genes differentially expressed between each CAF subtype in CRC **(Fig. 4a-d)**[43]. We also confirmed the presence of CAF-S1 and CAF-S4 signatures in pancreatic adenocarcinoma (n=118) and non-small cell lung cancer (NSCLC, n=80) cohorts (**Fig. 4a**) (see methods for datasets). Gene signatures were specific to each CAF-S1 and CAF-S4 in bulk transcriptomic datasets thus confirming their existence in TME of CRC and other tumor types.

We found high CAF-S1 and CAF-S4 signatures associated with significantly poor median overall-survival, irrespective of CMS in three independent CRC datasets **(Fig. 4b-d, Supplementary Fig. 6**). Additionally, CAF signatures stratified the CMS4 subtype into high- and low-risk overall survival in all datasets, thus identifying additional heterogeneity and providing prognostication in this aggressive patient subgroup (**Fig. 4b-d**). Here, using scRNA-seq, we show for the first time that high CAF infiltration in CRC is associated with poor prognosis across all molecular subtypes, and which further stratifies the CMS4 subgroup into high and low-risk clinical phenotypes in CRC cohorts.

Single-cell RNA sequencing reveals heterogeneity beyond Consensus Molecular Subtypes in colorectal cancers and offers therapeutic opportunities.

The lack of association between tumor epithelia and CMS classification, as well as the survival differences between high- and low-risk CAF signatures across CRC molecular subtypes suggest CRCs are less well defined than the traditional classification systems have indicated (e.g. those systems defined by somatic alterations, epigenomic features, and bulk gene expression data)[10–12,45,46].

To test our hypothesis, we estimated every cell type fraction using single-cell data in a discovery cohort (GSE39852[11], n=585) with a machine-learning algorithm, CIBERSORTx[47]. Cell type fractions were then validated in two independent cohorts of 177 (GSE17536[48]) and 290 (GSE14333[49]) tumors[9]. When we compared epithelial, immune, and stromal cell populations among the CMS subtypes, we did not detect a distinct pattern of tumor, immune, or stromal cell abundance across the four subtypes (**Fig. 5a-b, Supplementary Figs. 7a-b, 8a-b**). Although, we noted some trends in cell pattern enrichment amongst a few cell types consistent with CMS classification (CAFs were enriched in CMS4; dendritic cells, monocytes and TAMs were enriched in CMS1/CMS4; epithelial cells were enriched in CMS3) overall cellular phenotypes were present in varying proportions without a clear distinction between the four subtypes. The discovery and validation cohorts showed significant discordance in terms of cell phenotype enrichment with respect to each CMS subtype (**Fig. 5a-b, Supplementary Figs. 7a-b, 8a-b)**[9]. These discordant results could potentially be due to intra-tumoral variations in tumor purity, location of tumor biopsy, stromal and immune cell infiltration, and/or CMS's inability to address tumor-TME to tumor-TME and TME to TME variabilities, among other factors[9,50].

Based on the above findings, we hypothesized that CRC tumors may be more accurately represented as a continuum as has been proposed by Ma et al.[51]. The authors analyzed bulk transcriptomic data using a novel computational framework in which *denovo*, unsupervised clustering methods (k-medoid, non-negative-matrix factorization, and consensus clustering) best classified CRC tumors in a transcriptomic continuum[52–54]. They further carried out principal component analysis and robustly validated two principal components, PC Cluster Subtype Scores 1 and 2 (PCSS1 and PCSS2, respectively). Using this framework, we reasoned that single-cell data could elucidate the biological underpinnings of a CRC continuum model, and resolve stromal confounding seen using bulk transcriptomes[50,55].

We evaluated every cell fraction (epithelial, stromal, and immune components) using the Ma et al. algorithm on our discovery and validation cohorts, focusing on the validated PCSS1 and PCSS2. Upon projecting bulk transcriptomes onto the four CMS quadrants, we analyzed single cell-specific gene signatures after deconvolution. Intriguingly, we noted that each cell type not only projected in the expected quadrant but some of these same cell types also projected on other CMS quadrants **(Fig. 6a-b, supplementary 9a-b, 10a-b, Supplementary table 5-6)**. For example, epithelial cells, in addition to projecting on CMS2/CMS3 quadrants, also projected on other CMS quadrants and exhibited continuum. These data reflect significant intra-tumoral heterogeneity among all the cellular components that makeup the tumor ecosystem. Thus, it appears that CRC exists in a transcriptomic continuum not only with respect to the tumor cells themselves but also all the other cell types that make up the TME. These aspects would not have been apparent based on bulk transcriptomics analysis alone.

We found that transcriptional shifts were reproducible across discovery and validation datasets for major cell types (**Fig. 6a-b, supplementary 9a-b, 10a-b)**. Our analysis show no reliability in classifying CRC into immune–stromal rich (CMS1/CMS4) or immune-stromal desert (CMS2/CMS3) subtypes as proposed previously[16]. Thus, confirming continuous scores rather than discrete subtypes may improve classifying CRC tumors and may explain tumor-to-tumor variability, tumor/TME-to-tumor/TME and TME-to-TME variabilities within CMS subgroups as has been previously described [51]. While CAF-S1 and CAF-S4 were present across all CMS groups these CAFs exhibited high PCSS1 and PCSS2 scores correlating with increased enrichment in the CMS4 group. Thus, our analysis identified CAF-S1 and CAF-S4 as the cells of origin for biological heterogeneity in CMS4 subtypes associated with poor prognosis (**Supplementary table 5-6**).

DISCUSSION

In the present study, we evaluated the CMS classification system of CRC that has been established using bulk RNA-seq through the lens of single-cell transcriptomics. We identified significant intratumoral transcriptomic heterogeneity of cell states within tumor epithelial cells. These findings are consistent with a prior DNA-based multi-omics approach by Sottoriva et al. which proposed a ‘big bang model’ of CRC[56]. This model suggests that at initiation, CRC is composed of a mixture of subclones, underscoring the significant heterogeneity inherent in CRC biology and potentially reflecting non-linear pathways during CRC evolution and progression. Furthermore, both our findings (using scRNA-seq and pseudotime trajectory analysis) and Sottoriva et al. agree with clinical observations that tumors with diverse clonal populations respond unpredictably to monolithic treatment strategies “targeted” to average expression profiles based on bulk sequencing. Thus, targeting sub-clonal transcriptomic patterns is likely to be more effective when designing personalized CRC therapies [56] [57]. Interestingly, a recent study in gastroesophageal adenocarcinoma (GEA) that employed therapies to target dominant sub-clonal populations in metastatic GEA patients demonstrated improved survival as compared to standard monolithic treatment strategies[58].

We find that stromal cells are important contributors to the observed biological heterogeneity of CRC, in agreement with emerging data that suggest stromal-derived signatures play key role in modulating CRC biology[7–9,37,59,60]. These recent studies employing bulk transcriptomics demonstrated that the degree of stromal infiltration is associated with prognosis, while a small scRNAseq study utilizing 26 fibroblasts showed poor survival particularly among the CMS4 subtype CAF-enriched CRC tumors[61]. By using a much larger sample set of 1,182 high-quality fibroblasts, our current work identifies the CAF-S1 and CAF-S4 subtypes to be the cells of origin associated with poor prognosis across all CRC patients and not just those with CMS4 classified tumors[61]. CAF signatures are able to stratify patients by median overall survival (**Fig. 4b-d**). These findings are significant since the CMS4 subtype, is primarily stromal-driven and is enriched in more than 40% of metastatic CRC samples from patients with worse outcomes[17,18]. Thus, targeting CAFs to remodel the tumor microenvironment may lead to improved and much-needed therapeutic development for metastatic CRC patients[17,18].

Targeting of CAFs in solid tumors is being explored in multiple clinical trials with variable results[62,63]. Such studies likely failed to address CAF heterogeneity and their complex interactions with the other cells of TME. Our study suggests that CRC may be intricately entwined with the stroma, and therefore may be amenable to stromal targeted combinatoric approaches, including monoclonal antibodies that abrogate CAF-S1 function. In future studies, the treatment of CRC patients should involve stroma targeted therapies and take the above aspects into consideration[63,64]. The scRNA-seq or bulk-RNA-seq signatures corresponding to CAF-S1 and CAF-S4 may serve as suitable biomarkers for tumors that are reliant on this axis.

Immunotherapy responses in MSS CRC, which comprise almost 95% of metastatic CRC, are lacking[65]; CAF subpopulations within the TME may be suppressing immune responses in these tumors. Based on our analysis we speculate that targeting ecm-myCAF and TGFβ-myCAF subtypes (responsible for immunotherapy resistance in NSCLC and melanoma) via bispecific antibodies, vaccines, or even cell-based therapies, may enhance current checkpoint blockade strategies[44,63,64]. Functional validation and clinical studies will be required to confirm the clinical utility of targeting these CAF populations in CRC.

More importantly, our study's single-cell resolution enables us to investigate whether tumor cell transcriptomes, and by extension, biological phenotypes, are the primary determinant of CMS classification. Based on our findings, it appears that bulk analysis may have been confounded by varying degrees of tumor microenvironment population enrichment, and that tumor cells within each patient do not segregate into static phenotypes but rather exhibit considerable plasticity. Our single-cell analysis uncovered complex and mixed cellular phenotypes among each cell-specific subpopulation, which projected in a transcriptomic continuum across the CMS groups .These findings were further supported by our scRNA-seq CMS classification analysis that assigned each CRC sample to multiple CMS groups, thereby suggesting CMS heterogeneity within each CRC tumor (**Fig. 1d, Supplementary Table 1**)[16,66]. These findings may also explain why CMS-defined populations of tumors have not been readily observed in transcriptomic data from independent CRC cohorts[9,17]. Our results suggest that attempts to divide CRC phenotypes into the current discrete subtypes may undermine optimal patient stratification in the clinical trial setting[9]. The only prospective study to date that utilized the CMS classification (specifically the CMS4 subtype) for patient selection based on dual PD-L1/TGF-ß expression signatures was halted due to futility, suggesting CMS classification may adversely impact clinical trial design[67].

Previous studies relying on bulk transcriptomics had concerns that stromal content may conceal subtle critical gene signals originating from other key cellular phenotypes within the CRC spectrum, and thus affect CRC classification[8,9,50]. Our extensive single-cell analysis allowed for detailed evaluation of all cellular subtypes of CRC simultaneously, thereby uncovering contributions of the different components making up the complex cellular milieu within the CMS classification schema. The present analysis resolved the issue of stromal confounding in CRC and showed that all cellular components contribute to a transcriptomic continuum encompassing all the subtypes that together define the CMS system.

In conclusion, our analysis not only shows tumor-to-tumor variability (as proposed by Ma et al. and other groups), but also demonstrates tumor-TME to tumor-TME as well as TME to TME transcriptional variability at the single-cell resolution level[9,51]. These data contribute to the conceptual advances in understanding CRC pathogenesis, clinical management, and therapeutic development. We suggest revisiting CMS-like classifications that define CRCs into distinct static subtypes while, in reality CRC tumors exist in a continuum. We caution in using CMS to stratify patients for drug development and clinical trial design. Future studies should concentrate on developing biomarkers (such as CAF's) and therapeutic agents that can stratify CRC patients beyond traditional classification system boundaries. Ultimately, newer single-cell multiomic technologies will allow us to detect somatic mutations, transcriptomes, proteomes, epigenomes and metabolomes in real-time at the single-cell resolution to better guide more individualized and improved therapies.

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." We would also like to thank Dr. Kristian Pietras from providing differential gene expression list of CAFs from his paper titled "Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing". Above all, we want to thank our patients who participated in this study and their families.

AUTHOR CONTRIBUTIONS

A.M. devised, supervised the study, conducted data analysis, 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*.'

REFERENCES

1. Cancer of the Colon and Rectum - Cancer Stat Facts. SEER. Accessed December 29, 2020. https://seer.cancer.gov/statfacts/html/colorect.html
2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017;66(4):683-691. doi:10.1136/gutjnl-2015-310912
3. Kunst N, Alarid-Escudero F, Aas E, Coupé VMH, Schrag D, Kuntz KM. Estimating population-based recurrence rates of colorectal cancer over time in the United States. *Cancer Epidemiol Biomarkers Prev*. 2020;29(12):2710-2718. doi:10.1158/1055-9965.EPI-20-0490
4. Xie Y-H, Chen Y-X, Fang J-Y. Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduction and Targeted Therapy*. 2020;5(1):1-30. doi:10.1038/s41392-020-0116-z
5. Pagès F, Mlecnik B, Marliot F, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *The Lancet*. 2018;391(10135):2128-2139. doi:10.1016/S0140-6736(18)30789-X
6. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337. doi:10.1038/nature11252
7. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*. 2015;47(4):320-329. doi:10.1038/ng.3225
8. Isella C, Terrasi A, Bellomo SE, et al. Stromal contribution to the colorectal cancer transcriptome. *Nat Genet*. 2015;47(4):312-319. doi:10.1038/ng.3224
9. Dunne PD, McArt DG, Bradley CA, et al. Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential piagnostic value in colorectal cancer. *Clin Cancer Res*. 2016;22(16):4095-4104. doi:10.1158/1078-0432.CCR-16-0032
10. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*. 2013;19(5):619-625. doi:10.1038/nm.3175
11. Marisa L, Reyniès A de, Duval A, et al. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLOS Medicine*. 2013;10(5):e1001453. doi:10.1371/journal.pmed.1001453
12. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015;21(11):1350-1356. doi:10.1038/nm.3967
13. Lenz H-J, Ou F-S, Venook AP, et al. Impact of Consensus Molecular Subtype on Survival in Patients With Metastatic Colorectal Cancer: Results From CALGB/SWOG 80405 (Alliance). *J Clin Oncol*. 2019;37(22):1876-1885. doi:10.1200/JCO.18.02258
14. Stintzing S, Wirapati P, Lenz H-J, et al. Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. *Ann Oncol*. 2019;30(11):1796-1803. doi:10.1093/annonc/mdz387
15. Mooi JK, Wirapati P, Asher R, et al. The prognostic impact of consensus molecular subtypes (CMS) and its predictive effects for bevacizumab benefit in metastatic colorectal cancer: molecular analysis of the AGITG MAX clinical trial. *Ann Oncol*. 2018;29(11):2240-2246. doi:10.1093/annonc/mdy410
16. Sveen A, Cremolini C, Dienstmann R. Predictive modeling in colorectal cancer: time to move beyond consensus molecular subtypes. *Annals of Oncology*. 2019;30(11):1682-1685. doi:10.1093/annonc/mdz412
17. Fontana E, Eason K, Cervantes A, Salazar R, Sadanandam A. Context matters—consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals of Oncology*. 2019;30(4):520-527. doi:10.1093/annonc/mdz052

18. Khambata-Ford S, Garrett CR, Meropol NJ, et al. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J Clin Oncol*. 2007;25(22):3230-3237. doi:10.1200/JCO.2006.10.5437
19. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience*. 2018;7(7). doi:10.1093/gigascience/giy083
20. Zhang L, Li Z, Skrzypczynska KM, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell*. 2020;181(2):442-459.e29. doi:10.1016/j.cell.2020.03.048
21. Zhang L, Yu X, Zheng L, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*. 2018;564(7735):268-272. doi:10.1038/s41586-018-0694-x
22. Corridoni D, Antanaviciute A, Gupta T, et al. Single-cell atlas of colonic CD8 + T cells in ulcerative colitis. *Nature Medicine*. 2020;26(9):1480-1490. doi:10.1038/s41591-020-1003-4
23. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*. 2019;20(2):163-172. doi:10.1038/s41590-018-0276-y
24. Luoma AM, Suo S, Williams HL, et al. Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy. *Cell*. 2020;182(3):655-671.e22. doi:10.1016/j.cell.2020.06.001
25. Lee H-O, Hong Y, Etlioglu HE, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics*. 2020;52(6):594-603. doi:10.1038/s41588-020-0636-z
26. Mills JC, Sansom OJ. Reserve stem cells: Differentiated cells reprogram to fuel repair, metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci Signal*. 2015;8(385):re8. doi:10.1126/scisignal.aaa7540
27. Lambrechts D, Wauters E, Boeckx B, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med*. 2018;24(8):1277-1289. doi:10.1038/s41591-018-0096-5
28. Izar B, Tirosh I, Stover EH, et al. A single-cell landscape of high-grade serous ovarian cancer. *Nature Medicine*. 2020;26(8):1271-1279. doi:10.1038/s41591-020-0926-0
29. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004
30. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun*. 2019;10(1):2750. doi:10.1038/s41467-019-10737-5
31. Lim SB, Yeo T, Lee WD, et al. Addressing cellular heterogeneity in tumor and circulation for refined prognostication. *Proc Natl Acad Sci U S A*. 2019;116(36):17957-17962. doi:10.1073/pnas.1907904116
32. Wang R, Dang M, Harada K, et al. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. *Nat Med*. 2021;27(1):141-151. doi:10.1038/s41591-020-1125-8
33. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7. doi:10.1186/1471-2105-14-7
34. Sathe A, Grimes SM, Lau BT, et al. Single-cell genomic characterization reveals the cellular reprogramming of the gastric tumor microenvironment. *Clin Cancer Res*. 2020;26(11):2640-2653. doi:10.1158/1078-0432.CCR-19-3231
35. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381-386. doi:10.1038/nbt.2859
36. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*. 2017;14(10):979-982. doi:10.1038/nmeth.4402
37. Dalerba P, Kalisky T, Sahoo D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol*. 2011;29(12):1120-1127. doi:10.1038/nbt.2038

38. Vermeulen L, Todaro M, Mello F de S, et al. Single-cell cloning of colon cancer stem cells reveals a multi-lineage differentiation capacity. *PNAS*. 2008;105(36):13427-13432. doi:10.1073/pnas.0805706105

39. Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun*. 2020;11(1):2285. doi:10.1038/s41467-020-16164-1

40. Monk M, Holding C. Human embryonic genes re-expressed in cancer cells. *Oncogene*. 2001;20(56):8085-8091. doi:10.1038/sj.onc.1205088

41. Costa A, Kieffer Y, Scholer-Dahirel A, et al. Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell*. 2018;33(3):463-479.e10. doi:10.1016/j.ccell.2018.01.011

42. Wu SZ, Roden DL, Wang C, et al. Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *The EMBO Journal*. 2020;39(19). doi:10.15252/embj.2019104063

43. Bartoschek M, Oskolkov N, Bocci M, et al. Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. *Nat Commun*. 2018;9(1):5150. doi:10.1038/s41467-018-07582-3

44. Kieffer Y, Hocine HR, Gentric G, et al. Single-cell analysis reveals fibroblast clusters linked to immunotherapy resistance in cancer. *Cancer Discov*. 2020;10(9):1330-1351. doi:10.1158/2159-8290.CD-19-1384

45. Sadanandam A, Wang X, de Sousa E Melo F, et al. Reconciliation of classification systems defining molecular subtypes of colorectal cancer. *Cell Cycle*. 2014;13(3):353-357. doi:10.4161/cc.27769

46. De Sousa E Melo F, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med*. 2013;19(5):614-618. doi:10.1038/nm.3174

47. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*. 2019;37(7):773-782. doi:10.1038/s41587-019-0114-2

48. Jorissen RN, Gibbs P, Christie M, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin Cancer Res*. 2009;15(24):7642-7651. doi:10.1158/1078-0432.CCR-09-1431

49. Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*. 2010;138(3):958-968. doi:10.1053/j.gastro.2009.11.005

50. Dunne PD, Alderdice M, O'Reilly PG, et al. Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. *Nature Communications*. 2017;8(1):15657. doi:10.1038/ncomms15657

51. Ma S, Ogino S, Parsana P, et al. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biology*. 2018;19(1):142. doi:10.1186/s13059-018-1511-4

52. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Accessed January 22, 2021. https://link.springer.com/article/10.1007/s10852-005-9022-1

53. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572-1573. doi:10.1093/bioinformatics/btq170

54. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11(1):367. doi:10.1186/1471-2105-11-367

55. Isella C, Brundu F, Bellomo SE, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nature Communications*. 2017;8(1):15107. doi:10.1038/ncomms15107

56. Sottoriva A, Kang H, Ma Z, et al. A Big Bang model of human colorectal tumor growth. *Nature Genetics*. 2015;47(3):209-216. doi:10.1038/ng.3214

57. Loeb LA, Kohrn BF, Loubet-Senear KJ, et al. Extensive subclonal mutational diversity in human colorectal cancer and its significance. *PNAS*. 2019;116(52):26863-26872. doi:10.1073/pnas.1910301116

58. Catenacci DVT, Moya S, Lomnicki S, et al. Personalized Antibodies for Gastroesophageal Adenocarcinoma (PANGEA): A Phase II Study Evaluating an Individualized Treatment Strategy for Metastatic Disease. *Cancer Discov*. 2021;11(2):308-325. doi:10.1158/2159-8290.CD-20-1408

59. Tauriello DVF, Palomo-Ponce S, Stork D, et al. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature*. 2018;554(7693):538-543. doi:10.1038/nature25492

60. Calon A, Espinet E, Palomo-Ponce S, et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer Cell*. 2012;22(5):571-584. doi:10.1016/j.ccr.2012.08.013

61. Li H, Courtois ET, Sengupta D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*. 2017;49(5):708-718. doi:10.1038/ng.3818

62. Gascard P, Tlsty TD. Carcinoma-associated fibroblasts: orchestrating the composition of malignancy. *Genes Dev*. 2016;30(9):1002-1019. doi:10.1101/gad.279737.116

63. Sahai E, Astsaturov I, Cukierman E, et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer*. 2020;20(3):174-186. doi:10.1038/s41568-019-0238-1

64. Liu T, Han C, Wang S, et al. Cancer-associated fibroblasts: an emerging target of anti-cancer immunotherapy. *J Hematol Oncol*. 2019;12(1):86. doi:10.1186/s13045-019-0770-1

65. André T, Shiu K-K, Kim TW, et al. Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer. *N Engl J Med*. 2020;383(23):2207-2218. doi:10.1056/NEJMoa2017699

66. Laurent-Puig P, Marisa L, Ayadi M, et al. Colon cancer molecular subtype intratumoral heterogeneity and its prognostic impact: An extensive molecular analysis of the PETACC-8. *Annals of Oncology*. 2018;29:viii18. doi:10.1093/annonc/mdy269.058

67. Mehrvarz Sarshekeh A, Lam M, Zorrilla IR, et al. Consensus molecular subtype (CMS) as a novel integral biomarker in colorectal cancer: A phase II trial of bintrafusp alfa in CMS4 metastatic CRC. *JCO*. 2020;38(15_suppl):4084-4084. doi:10.1200/JCO.2020.38.15_suppl.4084

68. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

69. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst*. 2019;8(4):329-337.e4. doi:10.1016/j.cels.2019.03.003

70. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106

71. Thompson B. Canonical correlation analysis. In: *Encyclopedia of Statistics in Behavioral Science*. American Cancer Society; 2005. doi:10.1002/0470013192.bsa068

72. Jolliffe I. Principal component analysis. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Springer; 2011:1094-1096. doi:10.1007/978-3-642-04898-2_455

73. Maaten L van der. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*. 2014;15(93):3221-3245.

74. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software*. 2018;3(29):861. doi:10.21105/joss.00861

75. Zilionis R, Engblom C, Pfirschke C, et al. Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity*. 2019;50(5):1317-1334.e10. doi:10.1016/j.immuni.2019.03.009

76. Helmink BA, Reddy SM, Gao J, et al. B cells and tertiary lymphoid structures promote immunotherapy response. *Nature*. 2020;577(7791):549-555. doi:10.1038/s41586-019-1922-8

77. Guo X, Zhang Y, Zheng L, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med*. 2018;24(7):978-985. doi:10.1038/s41591-018-0045-3

78. Szabo PA, Levitin HM, Miron M, et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun*. 2019;10(1):4706. doi:10.1038/s41467-019-12464-3

79. Nirschl CJ, Suárez-Fariñas M, Izar B, et al. IFNγ-dependent tissue-Immune homeostasis Is co-opted in the tumor microenvironment. *Cell*. 2017;170(1):127-141.e15. doi:10.1016/j.cell.2017.06.016

80. Shi Z, Zhang Q, Yan H, et al. More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discov*. 2019;5:64. doi:10.1038/s41421-019-0137-3

81. Ramesh A, Schubert RD, Greenfield AL, et al. A pathogenic and clonally expanded B cell transcriptome in active multiple sclerosis. *Proc Natl Acad Sci U S A*. 2020;117(37):22932-22943. doi:10.1073/pnas.2008523117

82. Puram SV, Tirosh I, Parikh AS, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*. 2017;171(7):1611-1624.e24. doi:10.1016/j.cell.2017.10.044

83. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*. 2017;14(3):309-315. doi:10.1038/nmeth.4150

84. Beaubier N, Tell R, Lau D, et al. Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. *Oncotarget*. 2019;10(24):2384-2396. doi:10.18632/oncotarget.26797

85. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Institute; 2012. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

86. Jorissen RN, Lipton L, Gibbs P, et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res*. 2008;14(24):8061-8069. doi:10.1158/1078-0432.CCR-08-1431

87. Skrzypczak M, Goryca K, Rubel T, et al. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One*. 2010;5(10). doi:10.1371/journal.pone.0013091

88. Kemper K, Versloot M, Cameron K, et al. Mutations in the Ras-Raf Axis underlie the prognostic value of CD133 in colorectal cancer. *Clin Cancer Res*. 2012;18(11):3132-3141. doi:10.1158/1078-0432.CCR-11-3066

89. Schlicker A, Beran G, Chresta CM, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics*. 2012;5:66. doi:10.1186/1755-8794-5-66

90. Janky R, Binda MM, Allemeersch J, et al. Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma. *BMC Cancer*. 2016;16:632. doi:10.1186/s12885-016-2540-6

91. Meister M, Belousov A, EC X, et al. Intra-tumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer. *Journal of Bioinformatics Research Studies*. 2014;1.

92. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3):307-315. doi:10.1093/bioinformatics/btg405

93. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

94. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep*. 2017;7(1):16618. doi:10.1038/s41598-017-16747-x

95. Wickham H. *Ggplot2*. 2nd ed. Springer

96. Therneau T. *A Package for Survival Analysis in R*.; 2020.

97. Zhao X, Valen E, Parker BJ, Sandelin A. Systematic clustering of transcription start site landscapes. *PLOS ONE*. 2011;6(8):e23409. doi:10.1371/journal.pone.0023409

98. Bunis DG, Andrews J, Fragiadakis GK, Burt TD, Sirota M. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. *Bioinformatics*. Published online December 12, 2020. doi:10.1093/bioinformatics/btaa1011

99. *Morpheus: Interactive Heat Maps Using "morpheus.Js" and "Htmlwidgets."*; 2021. https://software.broadinstitute.org/morpheus

100. Pedersen TL. *Patchwork: The Composer of Plots*. *R*.; 2020. https://CRAN.R-project.org/package=patchwork

101. Kassambara A. *Ggpubr: "ggplot2" Based Publication Ready Plots*. Based Publication; 2020. https://CRAN.R-project.org/package=ggpubr

FIGURE LEGENDS

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

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

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 of fibroblasts colored by specific CAF-S1 subtypes. **b,** Heatmap showing the variable expression of fibroblast specific marker genes across CAF-S1, CAF-S4, and normal fibroblasts. **c,** UMAP color-coded for marker genes for five CAF-S1 subtypes as indicated.

Figure 4. Correlation of CAF-S1 and CAF-S4 gene profiles across human bulk transcriptomic data. a, Pearson's correlation of genes from CAF-S1 and CAF-S4 profiles in colorectal cancer (n= 1584; CAF-S1 and CAF-S4 r > 0.8), pancreatic cancer (n= 118; CAF-S1 r = 0.70; CAF-S4 r= 0.60), non-small cell lung cancer (n = 80; CAF-S1 r = 0.69; CAF-S4 r = 0.67). **b-d,** Pearson correlation plots, Kaplan-Meyer survival curves, and bar plots of CMS status assessing CAF expression in individual CRC datasets. Plots b-c are generated from single GEO datasets; GSE17536 (n = 177), GSE39582 (n= 585) and GSE33113 (n= 96), respectively. Note: High CAF-S1 and CAF-S4 gene signatures are associated with poor survival across all CMS . r = coefficient correlation. Hazard ratio > 1 in figures b-d.

Figure 5. Average cell type abundance from CRC dataset GSE39582 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 pairwise t-tests comparisons (with Benjamani-Hochberg correction) and ANOVA tests of cell abundance across CMS are shown in the figure. **b,** Deconvolution heatmap of different cell types by average expression using CIBERSORTx demonstrating cell type distribution within each CMS category. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.

Figure 6. Continuous scores for CRC dataset GSE39582. a, Principal component analysis plot showing PCSS1 and PCSS2 continuous scores reported by CMS classification across 19 cell types show minimal separation in the top 2 principal components. **b,** All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Note that the cell types largely form a continuum along CMS status and are not clustered in discrete quadrants separate from one another. Cells and markers are colored by bulk CMS status accordingly to the tumor sample of origin. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.

METHODS

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

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

scRNA-seq data quality control, gene-expression quantification, dimensionality reduction, and identification of cell clusters. *Cell Ranger* was utilized to process the raw gene expression matrices per samples and all samples from multiple patients were combined in R package (v3.6.3 2020-02-29] -- "*Holding the Windsock*"). Seurat package (v3.2.2) was used in this integrative multimodal analysis[68]. Genes detected in fewer than three cells and cells expressing less than 200 detected genes were filtered out and excluded from analysis. In addition, cells expressing > 25% mitochondria were removed. Cell cycle scoring was performed, (for the S phase and the G2M phase) and the predicted cell cycle phases were calculated. Doublet detection and any higher-order multiplets that were not dissociated during sample preparation were removed via the *DoubletFinder* (v2.0.2) package using default settings[69]. Following quality control one normal colon sample (B-cac13) was discarded due to poor data quality. Finally, 49,859 cells remained and were utilized for downstream analysis.

We adopted the general protocol described in Stuart et al. (2019) to group single cells into different cell subsets[68]. We employed the following steps: clustering the cells within each compartment (including the selection of variable genes for each dataset based on a variance stabilizing transformation [VST]), canonical correlation analysis (CCA) to remove batch effects among the samples, reduction of dimensionality, and projection of cells onto graphs [70,71]. Principal component analysis (PCA) was carried out on the scaled data of highly variable genes[72]. The first 30 principal components (PCs) were used to cluster the cells and to perform a subtype analysis by nonlinear dimensionality reduction (t-SNE) and to construct Uniform Manifold Approximation and Projection (UMAP) for cell embeddings[73,74]. We identified cell clusters under the optimal resolution by a shared nearest neighbor (SNN) modularity optimization-based clustering method. We implemented the *FindClusters* function of the Seurat package, which first calculated *k-nearest neighbors* and constructed the SNN graph. We implemented the original *Louvain algorithm* (algorithm = 1) for modularity optimization. Additionally, we utilized Clustree (v0.4.3) and manual review for identifying the best clustering resolution[19].

Major cell type detection and data visualization. To identify all major cell types, we evaluated differentially expressed markers in each identity cell group by comparing them to other clusters using the Seurat *FindAllMarkers* function. We used positively expressed genes with an average expression of >/= 2-fold higher in that subcluster than the average expression in the rest of the other subclusters. We used known marker genes, which have the highest fold expression in that cluster with respect to the other clusters. We also utilized SingleR ((v0.99.10, R Package), which leverage large transcriptomic datasets of well-annotated cell types and manual annotation for cell-type identification[27,75–77]. Depending on the presence of known marker genes the clusters were grouped as: epithelial cells (*EPCAM, KRT8,* and *KRT18*), fibroblasts (*COL1A1, DCN, COL1A2,* and *C1R*), endothelial cells (CD31+), myeloid cells (*LYZ, MARCO, CD68,* and *FCGR3A*), CD4 T cells (*CD4*), CD8 T cells (*CD8A* and *CD8B*), and B cells (*MZB1*), [27,39,43,75,78–81]. The cells were eventually assembled into DGE matrices within each compartment, containing all six cell types.

Major-cell type subclustering and data visualization. Each major cell type, including epithelial cells, endothelial cells, T cells, B cells, myeloid cells, and fibroblasts was reclustered and reanalyzed to study each compartment at a higher resolution to detect granular cellular heterogeneity in CRC. Clustree (v0.4.3) and manual review were utilized for optimal cluster detection. For cell annotation of each cell type, we utilized published literature gene expression signatures and manual review of differential genes among clusters. Additionally, we again utilized SingleR (v0.99.10, R Package) for unbiased cell annotation. Interestingly, reclustering of major compartments individually also detected clusters expressing hybrid markers as well as cell clusters expressing markers from distinct lineages (such as T cell clusters expressing B cells); these were removed and excluded for further analysis. We utilized UMAP for visualization purposes. For validation, we analyzed 65,362 cells from 23 patients and applied the similar quality control metrics as outlined above. In addition, we also applied *vars.to.regress* function to remove low quality cells in an unbiased manner. We retained 31,383 high-quality single cells for further analysis[25]. These high-quality cells were analyzed utilizing the same pipelines and parameters as that for our primary cohort (**Supplementary Figs. 4-5 and 11-12**).

The InferCNV (v1.2.1) package was used with default paramets to identify the evidence for somatic large-scale chromosomal copy number alteration in epithelial cells (*EPCAM+, KRT8+, KRT18+*)[82]. Normal epithelial cells were used as the control group.

Trajectory analysis. We used Monocle v.2 (v2.14.0), a reverse graph embedding method to reconstruct single-cell trajectories in tumor and normal epithelium[83]. In brief, we used UMI count matrices and the *negbinomial.size*() parameter to create a *CellDataSet* object in the default setting. We grouped projected cells on UMAP in default settings for visualization of monocle results. We defined the cumulative duration of the trajectory to show the average amount of transcriptional transition that a cell undergoes as it passes from the starting state to the end state. The cells were also ordered in pseudotime to explain the transition of cells from one state to another.

Pathway- Gene set variation analysis (GSVA). Pathway analysis was performed on the 50 hallmark gene sets downloaded from *Molecular Signatures Database (v7.2)*. We used GSVA (v1.34.0), a non-parametric, unsupervised method to estimate the gene set variations and evaluation of pathway enrichment, and pathway scores were calculated for each cell using standard settings [29,33].

DNA and bulk RNA library construction. DNA and bulk RNA sequencing was performed as previously described[84]. One hundred nanograms of DNA from each tumor was mechanically sheared to an average size of 200 bp. Using the *KAPA Hyper Prep Pack*, DNA libraries were packed, hybridized into the *xT probe* package, and amplified with the *KAPA HiFi HotStart*

ReadyMix. For uniformity, each sample needed to have 95% of all targeted base pairs sequenced to a minimum depth of 300x. One hundred nanograms of RNA per tumor sample was heat fragmented to a mean size of 200 base pairs in the presence of magnesium. Using random primers, the RNA was used for first-strand cDNA synthesis, followed by second-strand synthesis and A-tailing, adapter ligation, bead-based cleanup, and amplification of the library. After library planning, the *IDT xGEN Exome Test Panel* was hybridized with samples. Streptavidin-coated beads and target recovery were carried out, accompanied by amplification using the *KAPA HiFi* library amplification package. The RNA libraries were sequenced on an *Illumina HiSeq 4000* using patterned flow cell technology to achieve at least 50 million reads.

Detection of somatic variation on DNA sequencing data. The tumor and normal FASTQ files were paired. For quality management measurement, FASTQ files were evaluated using FASTQC and matched with Novoalign (Novocraft, Inc.)[84,85]. SAM files were generated and converted to BAM files. The BAM files were sorted, and duplicates were marked. Single nucleotide variations (SNVs) were called after alignment and sorting. For discovery of copy number alterations, the de-duplicated BAM files and the VCF generated from the variant calling pipeline were processed to computate read depth and variance of heterozygous germline SNVs between the tumor sample and normal sample. Binary circular segmentation was introduced and segments with strongly differential $\log_2$ ratios between the tumor and its comparator were chosen. From a combination of differential coverage in segmented regions and estimation of stromal admixture provided by analysis of heterozygous germline SNVs, an estimated integer copy number was determined

Microsatellite instability status. Probes for 43 microsatellite regions were developed using *Tempus xT* assay[84]. Tumors were categorized into three groups by the MSI classification algorithm as described by Tempus: microsatellite instability-high (MSI-H), microsatellite stable (MSS) or microsatellite equivocal (MSE). MSI screening for paired tumor-normal patients used reads mapped to the microsatellite loci with at least 5 bps flanking the microsatellite. The sample was graded as MSI-H if there was a >70% chance of MSI-H classification. If the likelihood of MSI-H status was 30-70%, the test findings were too ambiguous to interpret and those samples were listed as MSE. If there was a <30% chance of MSI-H status, the sample was called MSS. Additionally, IHC results were used to classify tumors into MSS or MSI molecular subtypes. Both of these modalities were concordant and produced the same results.

Bulk transcriptomics analysis. We downloaded and pooled eight colorectal gene expression datasets (GSE13067[86], GSE13294[86], GSE14333[48], GSE17536[49], GSE20916[87], GSE33113[88], GSE35896[89], and GSE39582[11]), a pancreatic cancer dataset (GSE62165[90]) and a non-small cell lung cancer dataset (GSE33532[91]) to validate our findings from the single cell compartments by deconvoluting the bulk gene expression profiles into pseudo single-cell resolutions. We used Affy (v1.64.0) for the data analysis and for exploration of Affymetrix oligonucleotide array probe level data[92]. Batch correction was carried out using the removeBatchEffect (v3.42.2) function of the LIMMA program and CMScaller for the CMS classification (see below)[93]. Three datasets (GSE17536[49], GSE33113[88], and GSE39582[11]) were utilized for clinical outcome analysis[93,94].

Correlation patterns in bulk gene expressions for CAF compartments. To identify the top correlated CAF-marker genes within the combined eight CRC datasets, three bulk CRC gene expression sets individually, pancreas cancer and lung cancer datasets. We first transformed the bulk gene expression sets with $\log_2$ transformation. Next, marker genes with an average $\log_2$ FC>/= 0.5 and p<0.05 obtained from the single cell data of CAF-S1 and CAF-S4 compartments were separately intersected with the bulk gene expression sets. Genes with an average Spearman correlation score greater than 0.8 were kept as the CAF signatures within the bulk gene expression. Heatmaps illustrating the correlation patterns within and between the CAF compartments were prepared with the heatmap.2 function from ggplot package (v3.1.1) utilizing the Pearson correlation coefficient. Heatmaps illustrating the correlation patterns within and between the CAF compartments were prepared using the ggplot package (v3.1.1) utilizing the Pearson correlation coefficient[95].

The Cox proportional hazard regression model was used to examine the significance of 20 cell types from scRNA-seq in bulk expression data. Each cell type's marker genes with an average logFC>1 and adjusted P<0.05 were intersected with the bulk expression datasets separately. We only kept the marker genes with a high correlation with each other in bulk, which provides an average correlation score of > 0.8. The average bulk expression of each cell type's remaining marker genes was calculated and used in the hazard regression model as the representative of this cell type. For analysis of relationships with patient outcome, univariate models were calculated using Cox proportional hazard regression (coxph function from survival R package)[96].

Deconvoluting public bulk gene expression profiles into pseudo single-cell expressions. We used CIBERSORTx v1 to estimate composition of various cell populations in GSE39582[11], GSE14333[48], and GSE17536[49] datasets[47]. Signature gene matrices were created using the expression profiles of 49,859 cells as the reference single cell profile. We ran the ‘hires’ module with default parameters except for the ‘rmbatchBmode,’ and the bulk-mode batch correction argument was set to true. After the deconvolution process, we normalized the gene expressions according to the cell fractions in each sample and calculated each gene’s Z-transformed expression values. The average normalized expression of each cell type across all samples was plotted with the heatmap.3 R function of the GMD package (v0.3.3)[97]. A signature matrix highlighting marker genes of the different cell types was prepared with a heatmap.2 R function of ggplot (v3.1.1).

Consensus molecular subtyping of colorectal cancer (CMS Classification). We used R package CMScaller(v0.9.2), a nearest template prediction (NTP) algorithm, for the classification of gene expression datasets[94]. We set the permutation number to 1000 to predict the CMS classes of the samples in the GEO datasets with a p-value < 0.05. We ran CMScaller with default parameters.

Continuous subtype discovery using scRNA-seq analysis. Bulk mRNA expression profiles of the GEO dataset (GSE39582[11]), composed of 585 samples in total, were deconvoluted into the pseudo single-cell expression profiles via CIBERSORTx utilizing the expression data consisting of 20 different well known cell types from our scRNA-seq dataset[47]. We transformed the deconvoluted expression matrix with log2 transformation. The principal components cluster subtype scores (PCSSs) of the CMS among the samples, were determined separately for each cell type using an algorithm published by Ma et al[51]. To obtain the PCSSs, the average loading vectors were used. The results obtained for 20 cell types were projected on the first two PCSSs (PCSS1 and PCSS2) as they were validated by Ma et al. in their analysis using 18 datasets. We also analyzed two datasets (GSE14333[48], GSE17536[49]) to independently confirm reproducibility of continuous scores.

Statistics and reproducibility. All statistical analyses and graphs were created in R (v3.6.3) and using a Python-based computational analysis tool. Schematic representations were made using the Inkscape (https://inkscape.org/) software. Dim plots, bar plots and box plots were generated using the dittoSeq (v1.1.7) package with default parameters[98]. Violin plots were generated using the patchwork (v1.1.0) package and ggplot2 (v3.3.2) package in R with default parameters. Heatmaps were generated using Morpheus.R with default parameters[99,100]. ANOVA and pair-wised t-tests for the CMS classes across the deconvoluted expression profiles were performed in R using the ggpubr R (v0.4.0) package[101]. The Box and Whisker plots were generated using the boxplot function of the R base package at default parameters. The mean of the $\log_2$ transformed deconvoluted expression value of the samples in each CMS group was demonstrated with a horizontal straight line within each box. The length of a boxplot corresponds to the interquartile range (IQR), which is defined as the range between the first and third quartiles (Q1 and Q3), whereas the whiskers are the upper and lower extreme values of the data (either data's extremum values, or the Q3+1.5*IQR and Q1-1.5*IQR values, whichever was less extreme). To test for differential score based on CIBERSORTx cell abundances between different molecular subtypes pairwise t-test with Benjamini-Hochberg correction was used.

Survival analysis. Survival curves were obtained according to the Kaplan-Meier method survfit (v3.2-7), and differences between survival distributions were assessed by Log-rank test. The patients were divided into two groups (high/poor and low/good risk) according to their median expression values (survminer (v0.4.8)). The surv_cutpoint function uses the maximally selected rank statistics and implements standard methods for the approximation of the null distribution of maximally selected rank statistics (maxstat (v0.7-25)).

The proportional hazard assumption was tested to examine the fit of the model for survival of the samples in three GEO datasets (GSE17536[49], GSE33113[88], and GSE39582[11]) with respect to the deconvoluted bulk mRNA expressions. For analysis of the relationships with patient outcome, multivariate models were calculated using the Cox proportional hazard regression (coxph survival R package)[96].

Figure 1

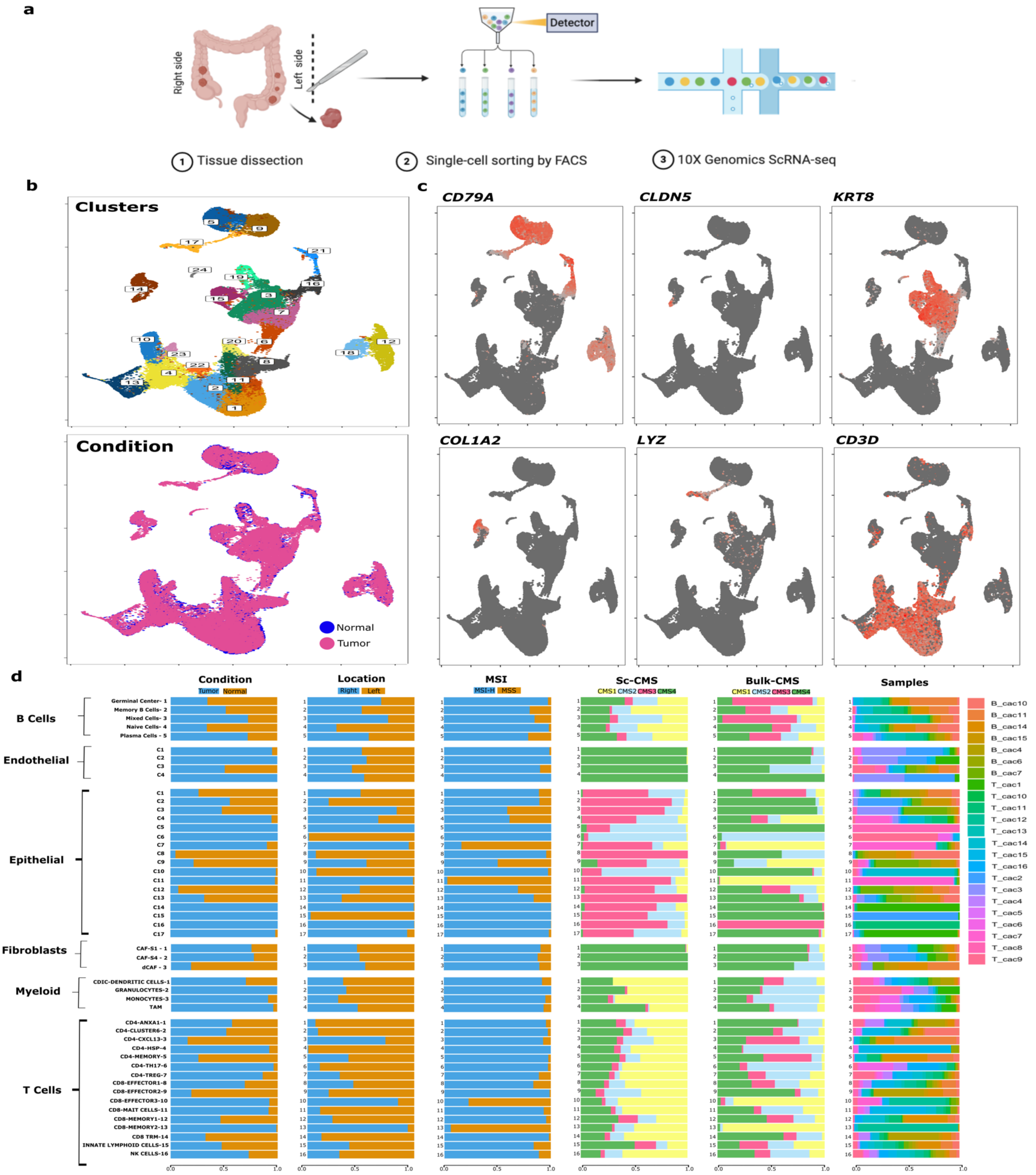

a
Right side
Left side
Detector
1 Tissue dissection
2 Single-cell sorting by FACS
3 10X Genomics ScRNA-seq
b
Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Condition
Normal
Tumor
c
CD79A
CLDN5
KRT8
COL1A2
LYZ
CD3D
d
Condition
Tumor Normal
Location
Right Left
MSI
MSI-H MSS
Sc-CMS
CMS1 CMS2 CMS3 CMS4
Bulk-CMS
CMS1 CMS2 CMS3 CMS4
Samples
B Cells
Germinal Center- 1
Memory B Cells- 2
Mixed Cells- 3
Naive Cells- 4
Plasma Cells - 5
Endothelial
C1
C2
C3
C4
Epithelial
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16
C17
Fibroblasts
CAF-S1 - 1
CAF-S4 - 2
dCAF - 3
Myeloid
CDIC-DENDRITIC CELLS-1
GRANULOCYTES-2
MONOCYTES-3
TAM
T Cells
CD4-ANXA1-1
CD4-CLUSTER6-2
CD4-CXCL13-3
CD4-HSP-4
CD4-MEMORY-5
CD4-TH17-6
CD4-TREG-7
CD8-EFFECTOR1-8
CD8-EFFECTOR2-9
CD8-EFFECTOR3-10
CD8-MAIT CELLS-11
CD8-MEMORY1-12
CD8-MEMORY2-13
CD8 TRM-14
INNATE LYMPHOID CELLS-15
NK CELLS-16
0.0
0.5
1.0
B_cac10
B_cac11
B_cac14
B_cac15
B_cac4
B_cac6
B_cac7
T_cac1
T_cac10
T_cac11
T_cac12
T_cac13
T_cac14
T_cac15
T_cac16
T_cac2
T_cac3
T_cac4
T_cac5
T_cac6
T_cac7
T_cac8
T_cac9

Figure 2

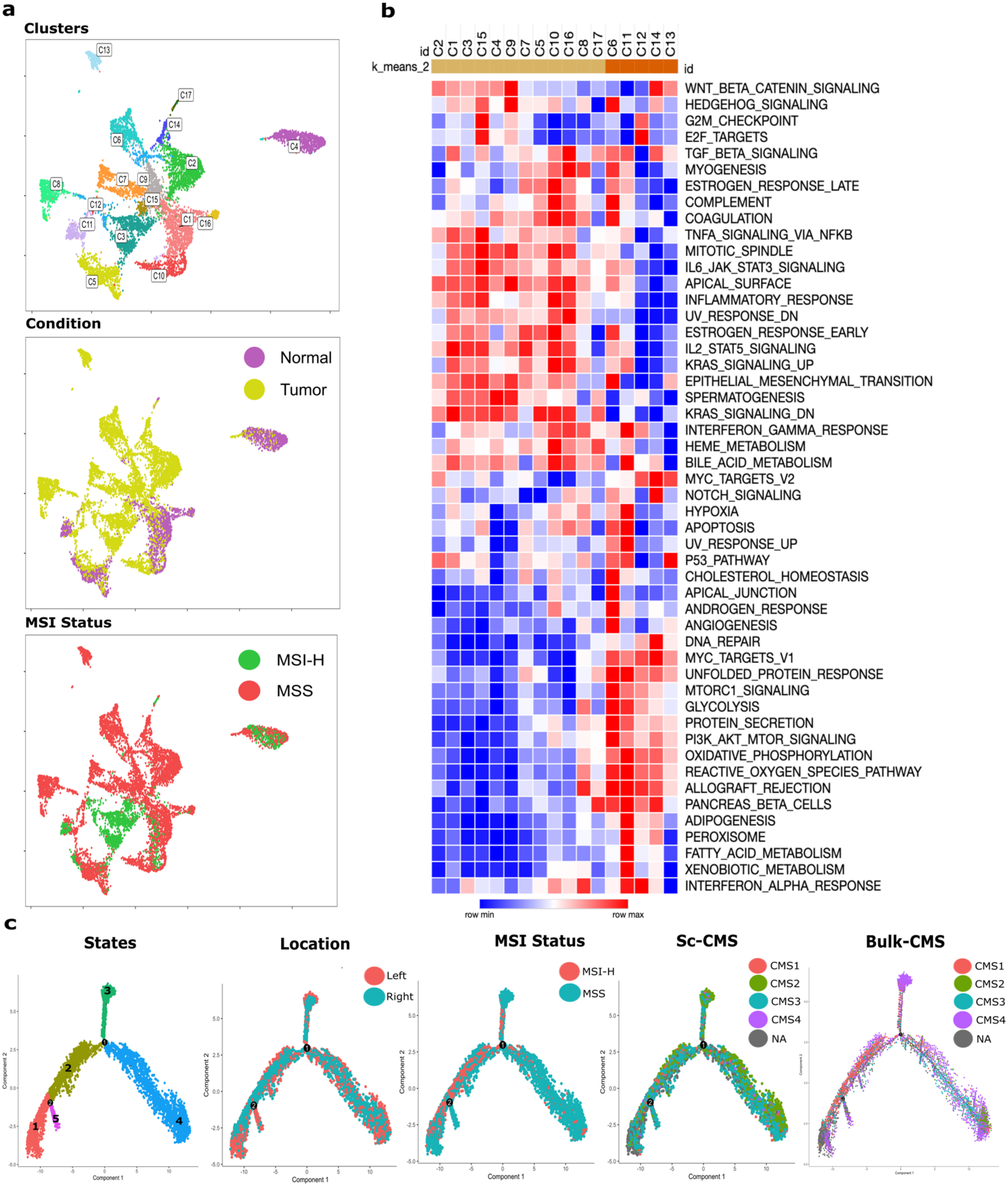

a
Clusters
C13
C17
C14
C4
C6
C2
C7
C9
C8
C15
C12
C11
C1
C16
C3
C5
C10
Condition
Normal
Tumor
MSI Status
MSI-H
MSS
b
id
C2
C1
C3
C15
C4
C9
C7
C5
C10
C16
C8
C17
C6
C11
C12
C14
C13
k_means_2
id
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
G2M_CHECKPOINT
E2F_TARGETS
TGF_BETA_SIGNALING
MYOGENESIS
ESTROGEN_RESPONSE_LATE
COMPLEMENT
COAGULATION
TNFA_SIGNALING_VIA_NFKB
MITOTIC_SPINDLE
IL6_JAK_STAT3_SIGNALING
APICAL_SURFACE
INFLAMMATORY_RESPONSE
UV_RESPONSE_DN
ESTROGEN_RESPONSE_EARLY
IL2_STAT5_SIGNALING
KRAS_SIGNALING_UP
EPITHELIAL_MESENCHYMAL_TRANSITION
SPERMATOGENESIS
KRAS_SIGNALING_DN
INTERFERON_GAMMA_RESPONSE
HEME_METABOLISM
BILE_ACID_METABOLISM
MYC_TARGETS_V2
NOTCH_SIGNALING
HYPOXIA
APOPTOSIS
UV_RESPONSE_UP
P53_PATHWAY
CHOLESTEROL_HOMEOSTASIS
APICAL_JUNCTION
ANDROGEN_RESPONSE
ANGIOGENESIS
DNA_REPAIR
MYC_TARGETS_V1
UNFOLDED_PROTEIN_RESPONSE
MTORC1_SIGNALING
GLYCOLYSIS
PROTEIN_SECRETION
PI3K_AKT_MTOR_SIGNALING
OXIDATIVE_PHOSPHORYLATION
REACTIVE_OXYGEN_SPECIES_PATHWAY
ALLOGRAFT_REJECTION
PANCREAS_BETA_CELLS
ADIPOGENESIS
PEROXISOME
FATTY_ACID_METABOLISM
XENOBIOTIC_METABOLISM
INTERFERON_ALPHA_RESPONSE
row min
row max
c
States
1
2
3
4
5
Location
Left
Right
MSI Status
MSI-H
MSS
Sc-CMS
CMS1
CMS2
CMS3
CMS4
NA
Bulk-CMS
CMS1
CMS2
CMS3
CMS4
NA
Component 1
Component 2
5.0
2.5
0.0
-2.5
-5.0
-10
-5
0
5
10

Figure 3

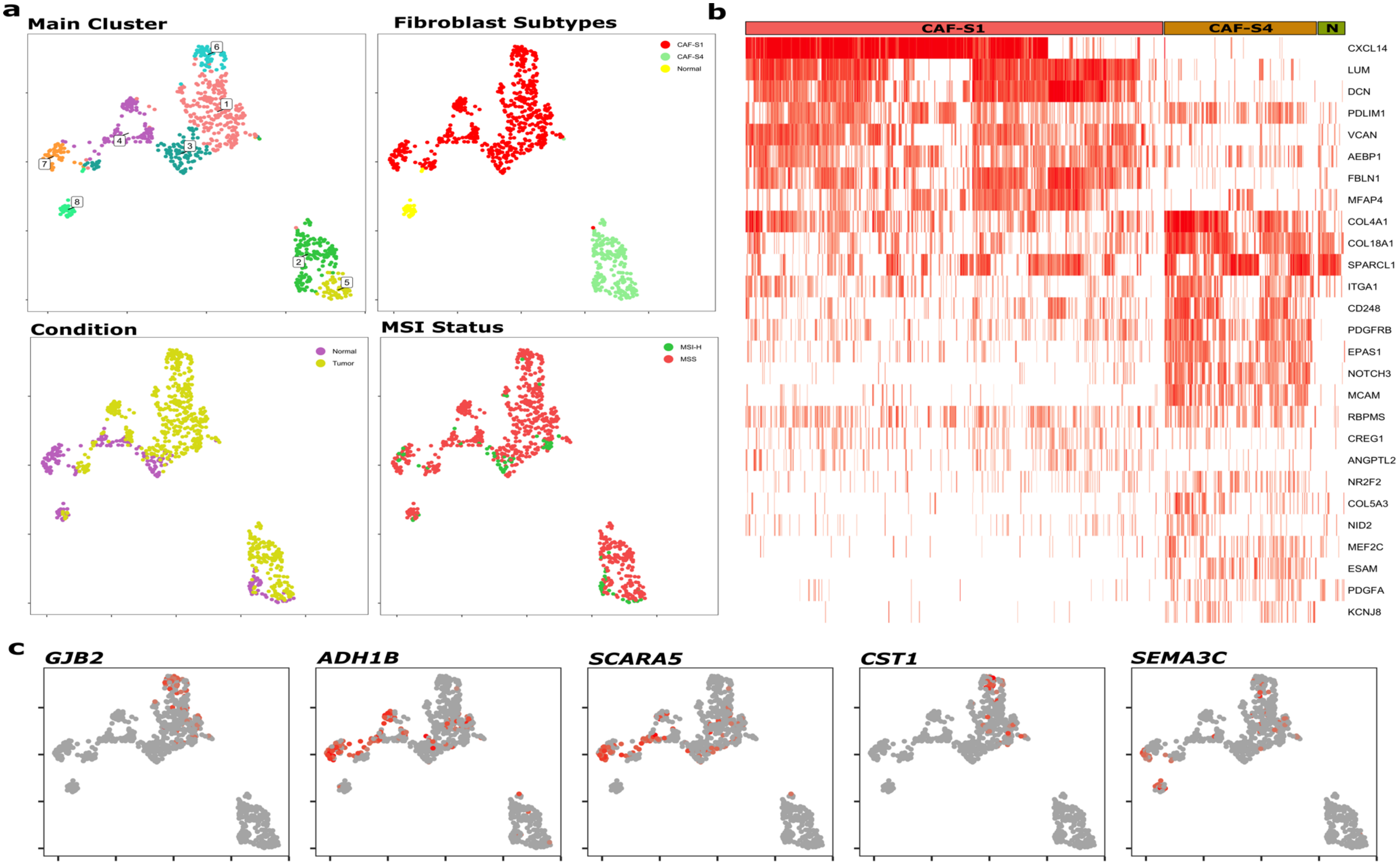

a
Main Cluster
1
2
3
4
5
6
7
8
Fibroblast Subtypes
CAF-S1
CAF-S4
Normal
Condition
Normal
Tumor
MSI Status
MSI-H
MSS
b
CAF-S1
CAF-S4
N
CXCL14
LUM
DCN
PDLIM1
VCAN
AEBP1
FBLN1
MFAP4
COL4A1
COL18A1
SPARCL1
ITGA1
CD248
PDGFRB
EPAS1
NOTCH3
MCAM
RBPMS
CREG1
ANGPTL2
NR2F2
COL5A3
NID2
MEF2C
ESAM
PDGFA
KCNJ8
c
GJB2
ADH1B
SCARA5
CST1
SEMA3C

Figure 4

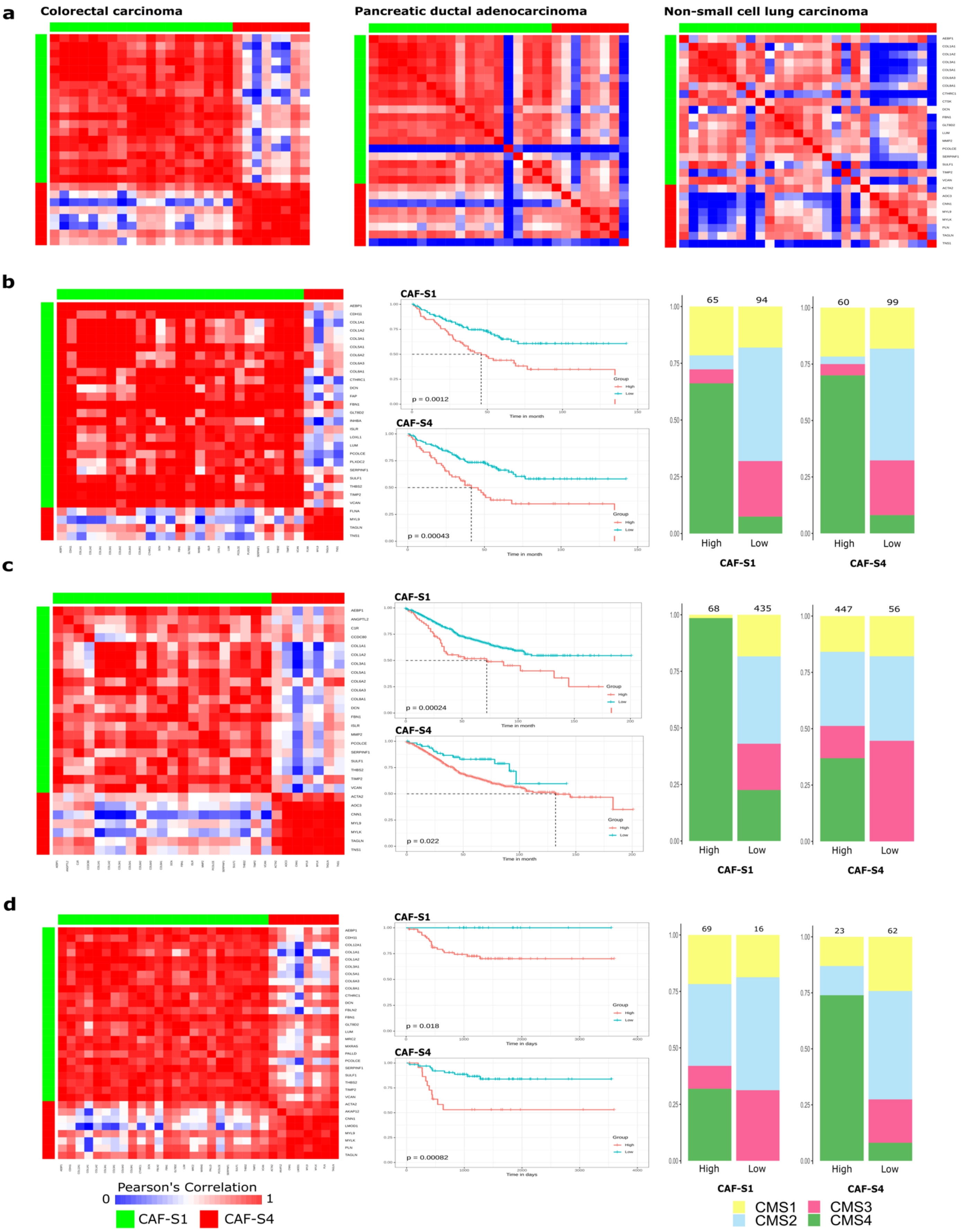

a
Colorectal carcinoma
Pancreatic ductal adenocarcinoma
Non-small cell lung carcinoma
AEBP1
COL1A1
COL1A2
COL3A1
COL5A1
COL6A3
COL8A1
CTHRC1
CTSK
DCN
FBN1
GLT8D2
LUM
MMP2
PCOLCE
SERPINF1
SULF1
TIMP2
VCAN
ACTA2
AOC3
CNN1
MYL9
MYLK
PLN
TAGLN
TNS1
b
AEBP1
CDH11
COL1A1
COL1A2
COL3A1
COL5A1
COL6A2
COL6A3
COL8A1
CTHRC1
DCN
FAP
FBN1
GLT8D2
INHBA
ISLR
LOXL1
LUM
PCOLCE
PLXDC2
SERPINF1
SULF1
THBS2
TIMP2
VCAN
FLNA
MYL9
TAGLN
TNS1
CAF-S1
p = 0.0012
Group
High
Low
Time in month
CAF-S4
p = 0.00043
65
94
60
99
High
Low
CAF-S1
CAF-S4
c
AEBP1
ANGPTL2
C1R
CCDC80
COL1A1
COL1A2
COL3A1
COL5A1
COL6A2
COL6A3
COL8A1
DCN
FBN1
ISLR
MMP2
PCOLCE
SERPINF1
SULF1
THBS2
TIMP2
VCAN
ACTA2
AOC3
CNN1
MYL9
MYLK
TAGLN
TNS1
CAF-S1
p = 0.00024
CAF-S4
p = 0.022
68
435
447
56
d
AEBP1
CDH11
COL12A1
COL1A1
COL1A2
COL3A1
COL5A1
COL6A3
COL8A1
CTHRC1
DCN
FBLN2
FBN1
GLT8D2
LUM
MRC2
MXRA5
PALLD
PCOLCE
SERPINF1
SULF1
THBS2
TIMP2
VCAN
ACTA2
AKAP12
CNN1
LMOD1
MYL9
MYLK
PLN
TAGLN
CAF-S1
p = 0.018
Time in days
CAF-S4
p = 0.00082
69
16
23
62
Pearson's Correlation
0
1
🟩 CAF-S1
🟥 CAF-S4
CMS1
CMS2
CMS3
CMS4

Figure 5

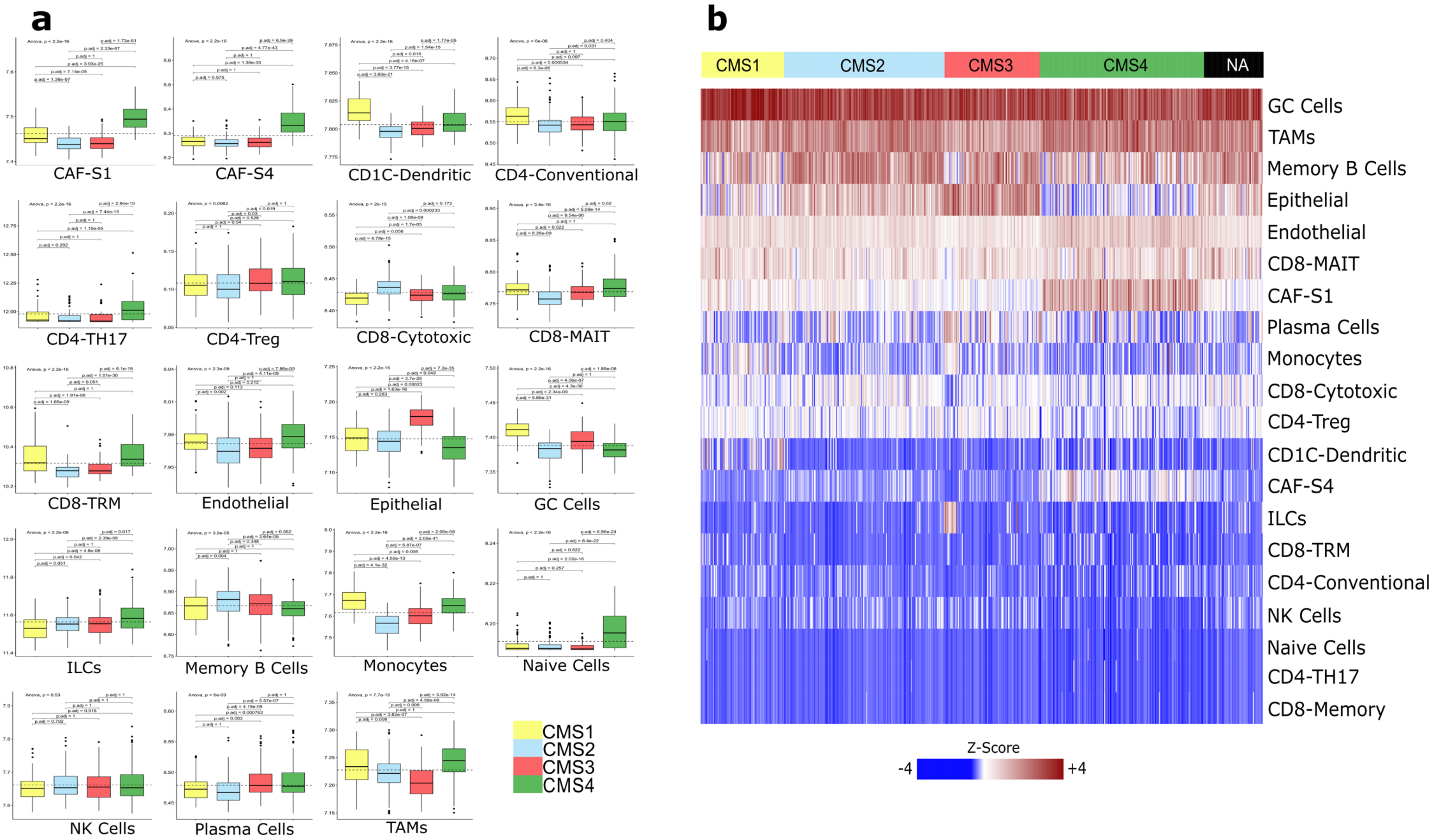
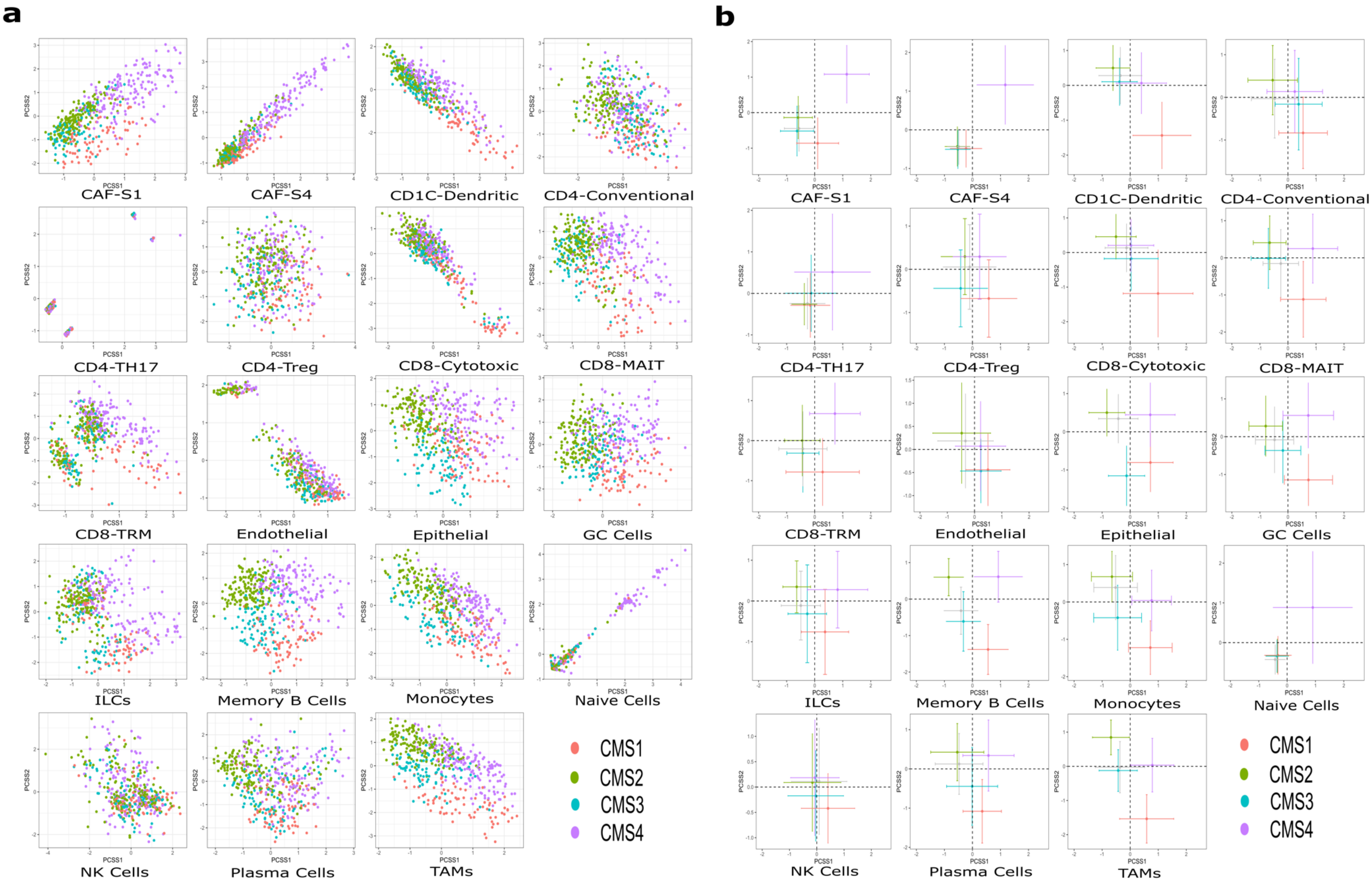

a
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-TH17
CD4-Treg
CD8-Cytotoxic
CD8-MAIT
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
CMS1
CMS2
CMS3
CMS4
b
CMS1
CMS2
CMS3
CMS4
NA
GC Cells
TAMs
Memory B Cells
Epithelial
Endothelial
CD8-MAIT
CAF-S1
Plasma Cells
Monocytes
CD8-Cytotoxic
CD4-Treg
CD1C-Dendritic
CAF-S4
ILCs
CD8-TRM
CD4-Conventional
NK Cells
Naive Cells
CD4-TH17
CD8-Memory
Z-Score
-4
+4

Figure 6

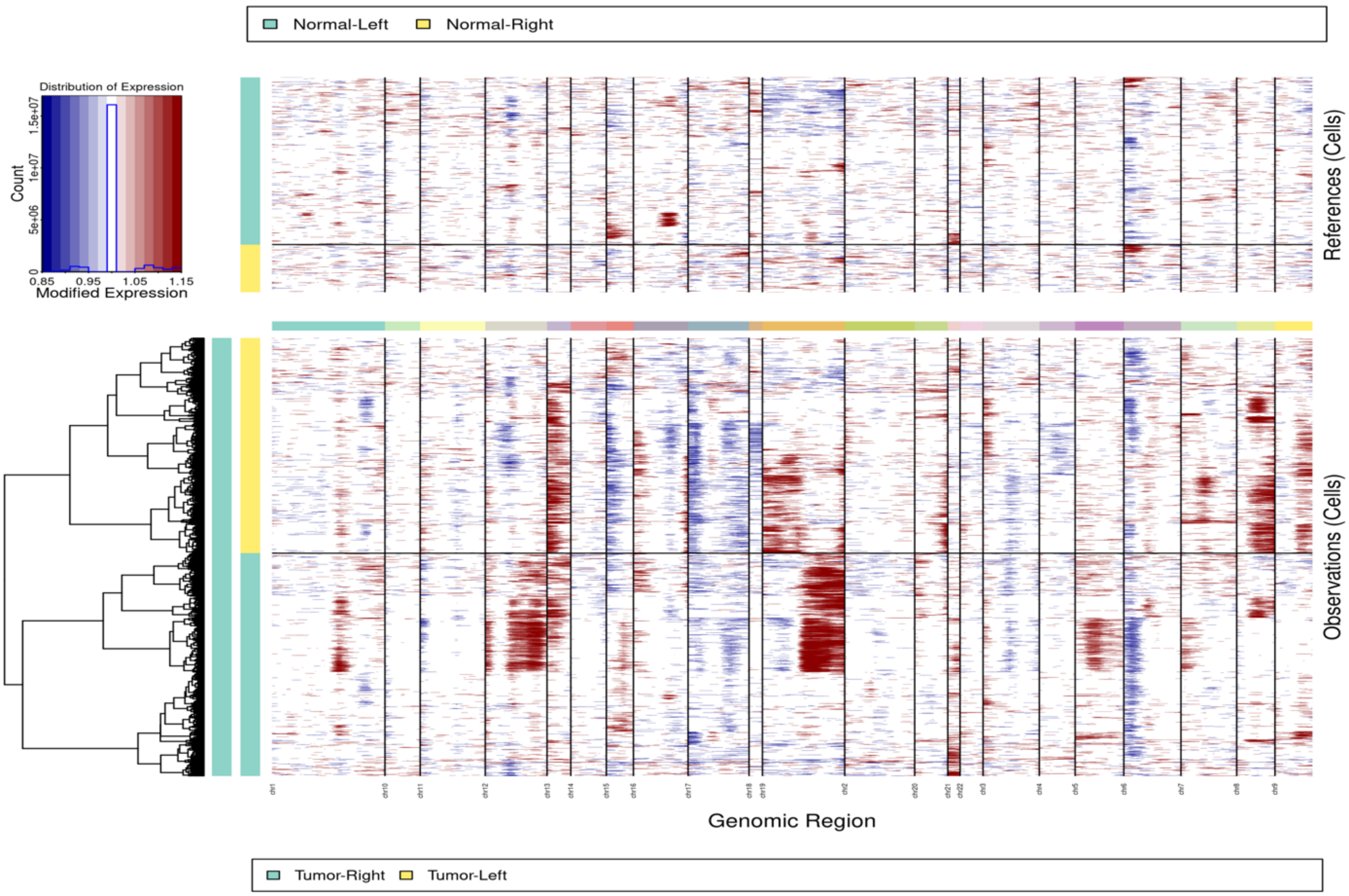

a
PCSS1
PCSS2
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-TH17
CD4-Treg
CD8-Cytotoxic
CD8-MAIT
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
CMS1
CMS2
CMS3
CMS4
b
PCSS1
PCSS2
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-TH17
CD4-Treg
CD8-Cytotoxic
CD8-MAIT
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
CMS1
CMS2
CMS3
CMS4

Supplementary Materials

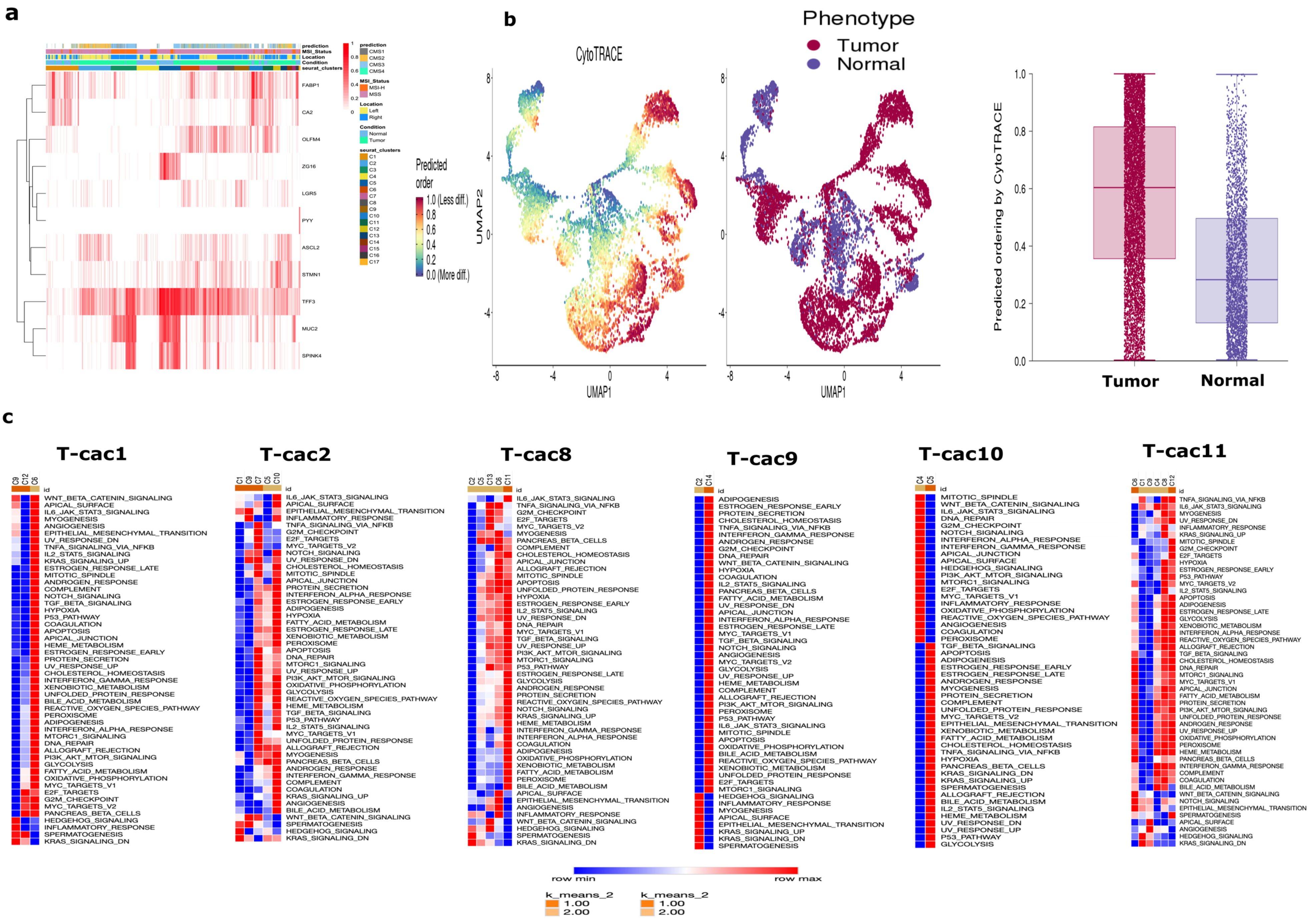
Khaliq et al.

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

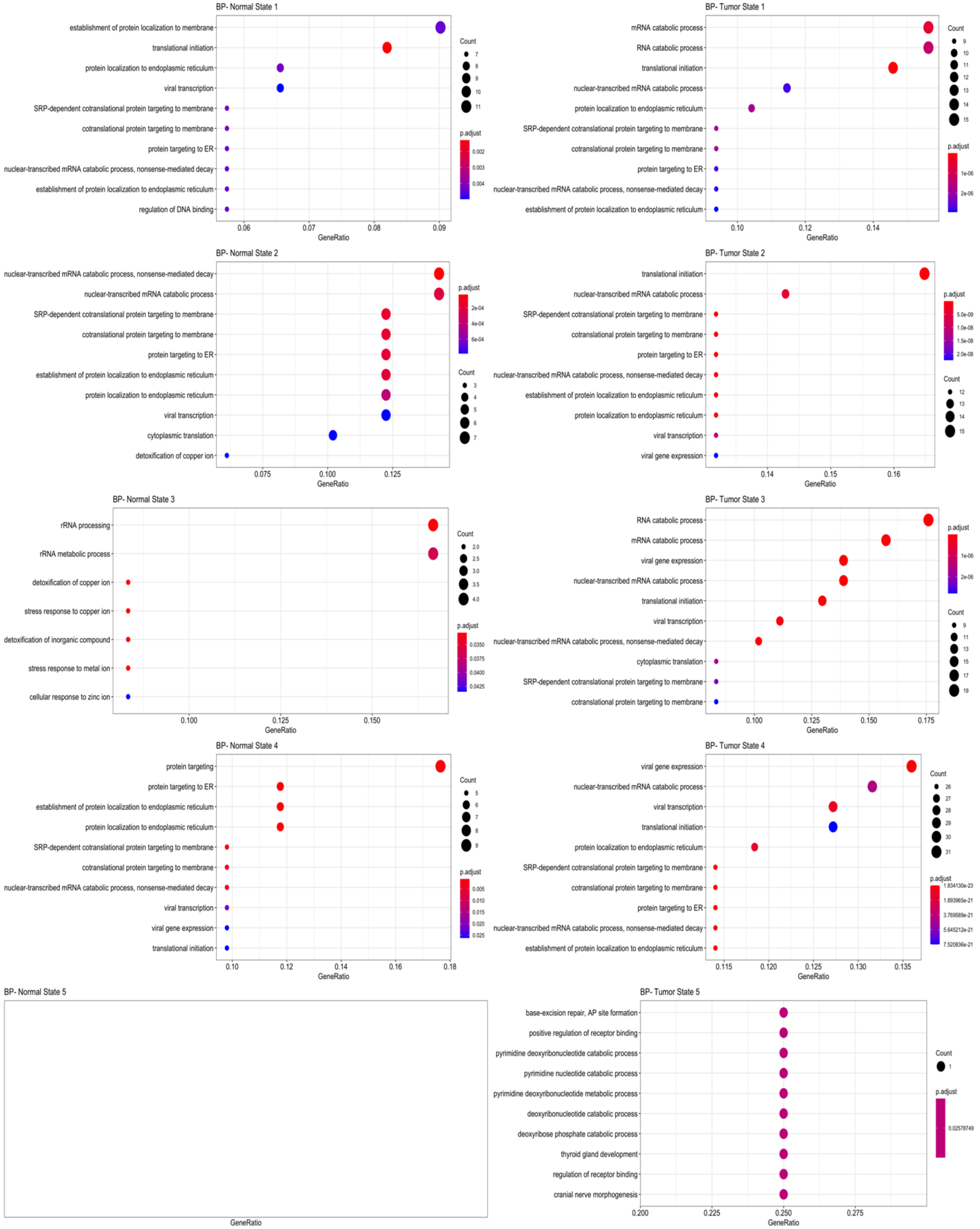
Supplementary figures 1-12



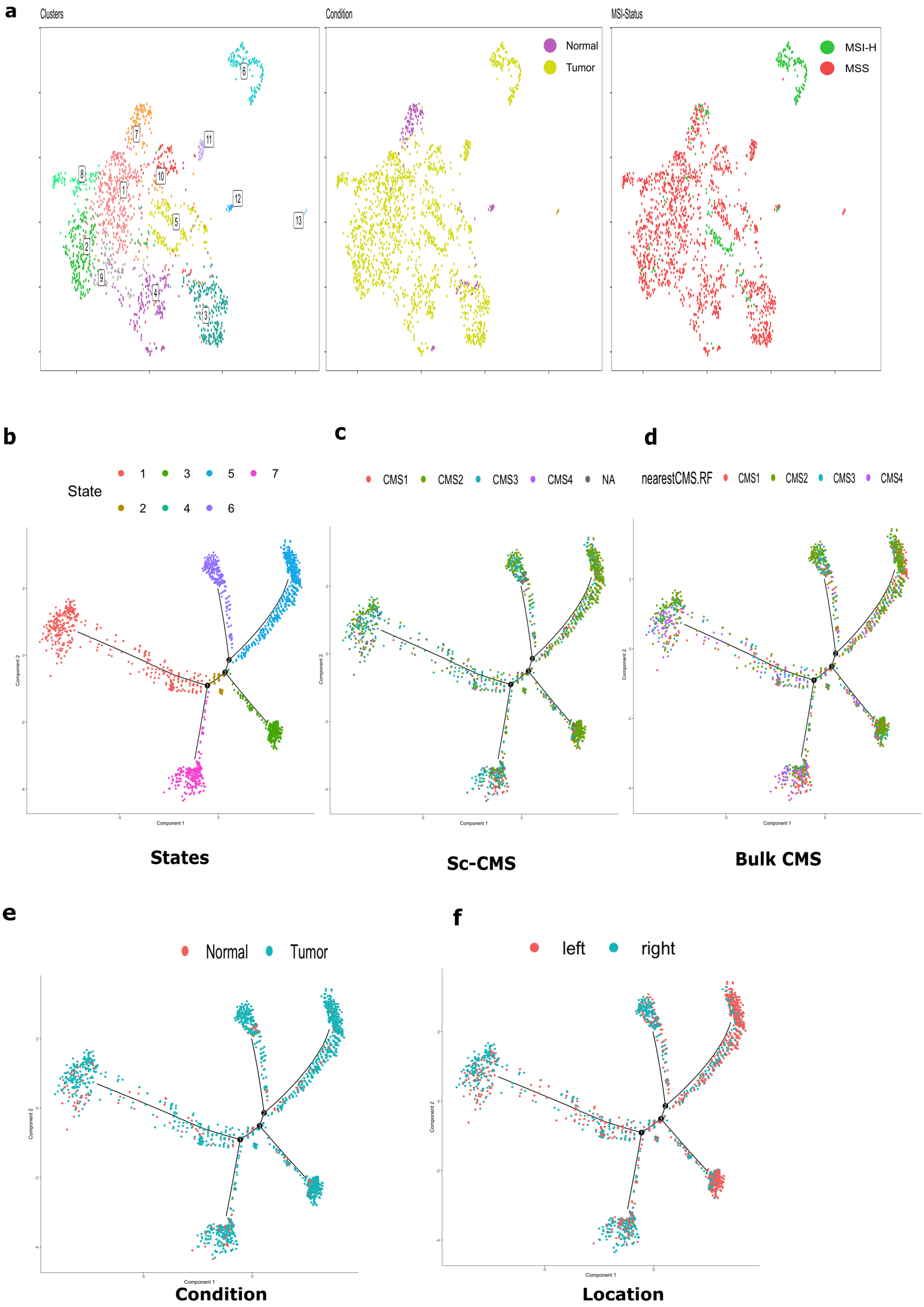
Supplementary figure 1. Analysis of copy number variation amongst epithelial cells. CNV analysis was conducted on the epithelial cell compartment. Increase CNV is seen in tumor derived epithelial cells (observation). Non-malignant derived epithelial cells were used as control (references).


a
prediction
MSI_Status
Location
Condition
seurat_clusters
FABP1
CA2
OLFM4
ZG16
LGR5
PYY
ASCL2
STMN1
TFF3
MUC2
SPINK4
1
0.8
0.6
0.4
0.2
0
prediction
CMS1
CMS2
CMS3
CMS4
MSI_Status
MSI-H
MSS
Location
Left
Right
Condition
Normal
Tumor
seurat_clusters
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16
C17
Predicted order
1.0 (Less diff.)
0.8
0.6
0.4
0.2
0.0 (More diff.)
b
CytoTRACE
Phenotype
Tumor
Normal
UMAP2
UMAP1
8
4
0
-4
-8
-4
0
4
Predicted ordering by CytoTRACE
1.0
0.8
0.6
0.4
0.2
0.0
Tumor
Normal
c
T-cac1
C9
C12
C6
id
WNT_BETA_CATENIN_SIGNALING
APICAL_SURFACE
IL6_JAK_STAT3_SIGNALING
MYOGENESIS
ANGIOGENESIS
EPITHELIAL_MESENCHYMAL_TRANSITION
UV_RESPONSE_DN
TNFA_SIGNALING_VIA_NFKB
IL2_STAT5_SIGNALING
KRAS_SIGNALING_UP
ESTROGEN_RESPONSE_LATE
MITOTIC_SPINDLE
ANDROGEN_RESPONSE
COMPLEMENT
NOTCH_SIGNALING
TGF_BETA_SIGNALING
HYPOXIA
P53_PATHWAY
COAGULATION
APOPTOSIS
APICAL_JUNCTION
HEME_METABOLISM
ESTROGEN_RESPONSE_EARLY
PROTEIN_SECRETION
UV_RESPONSE_UP
CHOLESTEROL_HOMEOSTASIS
INTERFERON_GAMMA_RESPONSE
XENOBIOTIC_METABOLISM
UNFOLDED_PROTEIN_RESPONSE
BILE_ACID_METABOLISM
REACTIVE_OXYGEN_SPECIES_PATHWAY
PEROXISOME
ADIPOGENESIS
INTERFERON_ALPHA_RESPONSE
MTORC1_SIGNALING
DNA_REPAIR
ALLOGRAFT_REJECTION
PI3K_AKT_MTOR_SIGNALING
GLYCOLYSIS
FATTY_ACID_METABOLISM
OXIDATIVE_PHOSPHORYLATION
MYC_TARGETS_V1
E2F_TARGETS
G2M_CHECKPOINT
MYC_TARGETS_V2
PANCREAS_BETA_CELLS
HEDGEHOG_SIGNALING
INFLAMMATORY_RESPONSE
SPERMATOGENESIS
KRAS_SIGNALING_DN
T-cac2
C1
C9
C7
C5
C10
id
IL6_JAK_STAT3_SIGNALING
APICAL_SURFACE
EPITHELIAL_MESENCHYMAL_TRANSITION
INFLAMMATORY_RESPONSE
TNFA_SIGNALING_VIA_NFKB
G2M_CHECKPOINT
E2F_TARGETS
MYC_TARGETS_V2
NOTCH_SIGNALING
UV_RESPONSE_DN
CHOLESTEROL_HOMEOSTASIS
MITOTIC_SPINDLE
APICAL_JUNCTION
PROTEIN_SECRETION
INTERFERON_ALPHA_RESPONSE
ESTROGEN_RESPONSE_EARLY
ADIPOGENESIS
HYPOXIA
FATTY_ACID_METABOLISM
ESTROGEN_RESPONSE_LATE
XENOBIOTIC_METABOLISM
PEROXISOME
APOPTOSIS
DNA_REPAIR
MTORC1_SIGNALING
UV_RESPONSE_UP
PI3K_AKT_MTOR_SIGNALING
OXIDATIVE_PHOSPHORYLATION
GLYCOLYSIS
REACTIVE_OXYGEN_SPECIES_PATHWAY
HEME_METABOLISM
TGF_BETA_SIGNALING
P53_PATHWAY
IL2_STAT5_SIGNALING
MYC_TARGETS_V1
UNFOLDED_PROTEIN_RESPONSE
ALLOGRAFT_REJECTION
MYOGENESIS
PANCREAS_BETA_CELLS
ANDROGEN_RESPONSE
INTERFERON_GAMMA_RESPONSE
COMPLEMENT
COAGULATION
KRAS_SIGNALING_UP
ANGIOGENESIS
BILE_ACID_METABOLISM
WNT_BETA_CATENIN_SIGNALING
SPERMATOGENESIS
HEDGEHOG_SIGNALING
KRAS_SIGNALING_DN
T-cac8
C2
C5
C13
C6
C11
id
IL6_JAK_STAT3_SIGNALING
TNFA_SIGNALING_VIA_NFKB
G2M_CHECKPOINT
E2F_TARGETS
MYC_TARGETS_V2
MYOGENESIS
PANCREAS_BETA_CELLS
COMPLEMENT
CHOLESTEROL_HOMEOSTASIS
APICAL_JUNCTION
ALLOGRAFT_REJECTION
MITOTIC_SPINDLE
APOPTOSIS
UNFOLDED_PROTEIN_RESPONSE
HYPOXIA
ESTROGEN_RESPONSE_EARLY
IL2_STAT5_SIGNALING
UV_RESPONSE_DN
DNA_REPAIR
MYC_TARGETS_V1
TGF_BETA_SIGNALING
UV_RESPONSE_UP
PI3K_AKT_MTOR_SIGNALING
MTORC1_SIGNALING
P53_PATHWAY
ESTROGEN_RESPONSE_LATE
GLYCOLYSIS
ANDROGEN_RESPONSE
PROTEIN_SECRETION
REACTIVE_OXYGEN_SPECIES_PATHWAY
NOTCH_SIGNALING
KRAS_SIGNALING_UP
HEME_METABOLISM
INTERFERON_GAMMA_RESPONSE
INTERFERON_ALPHA_RESPONSE
COAGULATION
ADIPOGENESIS
OXIDATIVE_PHOSPHORYLATION
XENOBIOTIC_METABOLISM
FATTY_ACID_METABOLISM
PEROXISOME
BILE_ACID_METABOLISM
APICAL_SURFACE
EPITHELIAL_MESENCHYMAL_TRANSITION
ANGIOGENESIS
INFLAMMATORY_RESPONSE
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
SPERMATOGENESIS
KRAS_SIGNALING_DN
T-cac9
C2
C14
id
ADIPOGENESIS
ESTROGEN_RESPONSE_EARLY
PROTEIN_SECRETION
CHOLESTEROL_HOMEOSTASIS
TNFA_SIGNALING_VIA_NFKB
INTERFERON_GAMMA_RESPONSE
ANDROGEN_RESPONSE
G2M_CHECKPOINT
DNA_REPAIR
WNT_BETA_CATENIN_SIGNALING
HYPOXIA
COAGULATION
IL2_STAT5_SIGNALING
PANCREAS_BETA_CELLS
FATTY_ACID_METABOLISM
UV_RESPONSE_DN
APICAL_JUNCTION
INTERFERON_ALPHA_RESPONSE
ESTROGEN_RESPONSE_LATE
MYC_TARGETS_V1
TGF_BETA_SIGNALING
NOTCH_SIGNALING
ANGIOGENESIS
MYC_TARGETS_V2
GLYCOLYSIS
UV_RESPONSE_UP
HEME_METABOLISM
COMPLEMENT
ALLOGRAFT_REJECTION
PI3K_AKT_MTOR_SIGNALING
PEROXISOME
P53_PATHWAY
IL6_JAK_STAT3_SIGNALING
MITOTIC_SPINDLE
APOPTOSIS
OXIDATIVE_PHOSPHORYLATION
BILE_ACID_METABOLISM
REACTIVE_OXYGEN_SPECIES_PATHWAY
XENOBIOTIC_METABOLISM
UNFOLDED_PROTEIN_RESPONSE
E2F_TARGETS
MTORC1_SIGNALING
HEDGEHOG_SIGNALING
INFLAMMATORY_RESPONSE
MYOGENESIS
APICAL_SURFACE
EPITHELIAL_MESENCHYMAL_TRANSITION
KRAS_SIGNALING_UP
KRAS_SIGNALING_DN
SPERMATOGENESIS
T-cac10
C4
C5
id
MITOTIC_SPINDLE
WNT_BETA_CATENIN_SIGNALING
IL6_JAK_STAT3_SIGNALING
DNA_REPAIR
G2M_CHECKPOINT
NOTCH_SIGNALING
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
APICAL_JUNCTION
APICAL_SURFACE
HEDGEHOG_SIGNALING
PI3K_AKT_MTOR_SIGNALING
MTORC1_SIGNALING
E2F_TARGETS
MYC_TARGETS_V1
INFLAMMATORY_RESPONSE
OXIDATIVE_PHOSPHORYLATION
REACTIVE_OXYGEN_SPECIES_PATHWAY
ANGIOGENESIS
COAGULATION
PEROXISOME
TGF_BETA_SIGNALING
APOPTOSIS
ADIPOGENESIS
ESTROGEN_RESPONSE_EARLY
ESTROGEN_RESPONSE_LATE
ANDROGEN_RESPONSE
MYOGENESIS
PROTEIN_SECRETION
COMPLEMENT
UNFOLDED_PROTEIN_RESPONSE
MYC_TARGETS_V2
EPITHELIAL_MESENCHYMAL_TRANSITION
XENOBIOTIC_METABOLISM
FATTY_ACID_METABOLISM
CHOLESTEROL_HOMEOSTASIS
TNFA_SIGNALING_VIA_NFKB
HYPOXIA
PANCREAS_BETA_CELLS
KRAS_SIGNALING_DN
KRAS_SIGNALING_UP
SPERMATOGENESIS
ALLOGRAFT_REJECTION
BILE_ACID_METABOLISM
IL2_STAT5_SIGNALING
HEME_METABOLISM
UV_RESPONSE_DN
UV_RESPONSE_UP
P53_PATHWAY
GLYCOLYSIS
T-cac11
C6
C1
C9
C4
C8
C12
id
TNFA_SIGNALING_VIA_NFKB
IL6_JAK_STAT3_SIGNALING
MYOGENESIS
UV_RESPONSE_DN
INFLAMMATORY_RESPONSE
KRAS_SIGNALING_UP
MITOTIC_SPINDLE
G2M_CHECKPOINT
E2F_TARGETS
HYPOXIA
ESTROGEN_RESPONSE_EARLY
P53_PATHWAY
MYC_TARGETS_V2
IL2_STAT5_SIGNALING
APOPTOSIS
ADIPOGENESIS
ESTROGEN_RESPONSE_LATE
GLYCOLYSIS
XENOBIOTIC_METABOLISM
INTERFERON_ALPHA_RESPONSE
REACTIVE_OXYGEN_SPECIES_PATHWAY
ALLOGRAFT_REJECTION
TGF_BETA_SIGNALING
CHOLESTEROL_HOMEOSTASIS
DNA_REPAIR
MTORC1_SIGNALING
MYC_TARGETS_V1
APICAL_JUNCTION
FATTY_ACID_METABOLISM
PROTEIN_SECRETION
PI3K_AKT_MTOR_SIGNALING
UNFOLDED_PROTEIN_RESPONSE
ANDROGEN_RESPONSE
UV_RESPONSE_UP
OXIDATIVE_PHOSPHORYLATION
PEROXISOME
HEME_METABOLISM
PANCREAS_BETA_CELLS
INTERFERON_GAMMA_RESPONSE
COMPLEMENT
COAGULATION
BILE_ACID_METABOLISM
WNT_BETA_CATENIN_SIGNALING
NOTCH_SIGNALING
EPITHELIAL_MESENCHYMAL_TRANSITION
SPERMATOGENESIS
APICAL_SURFACE
ANGIOGENESIS
HEDGEHOG_SIGNALING
KRAS_SIGNALING_DN
row min
row max
k_means_2
1.00
2.00
k_means_2
1.00
2.00

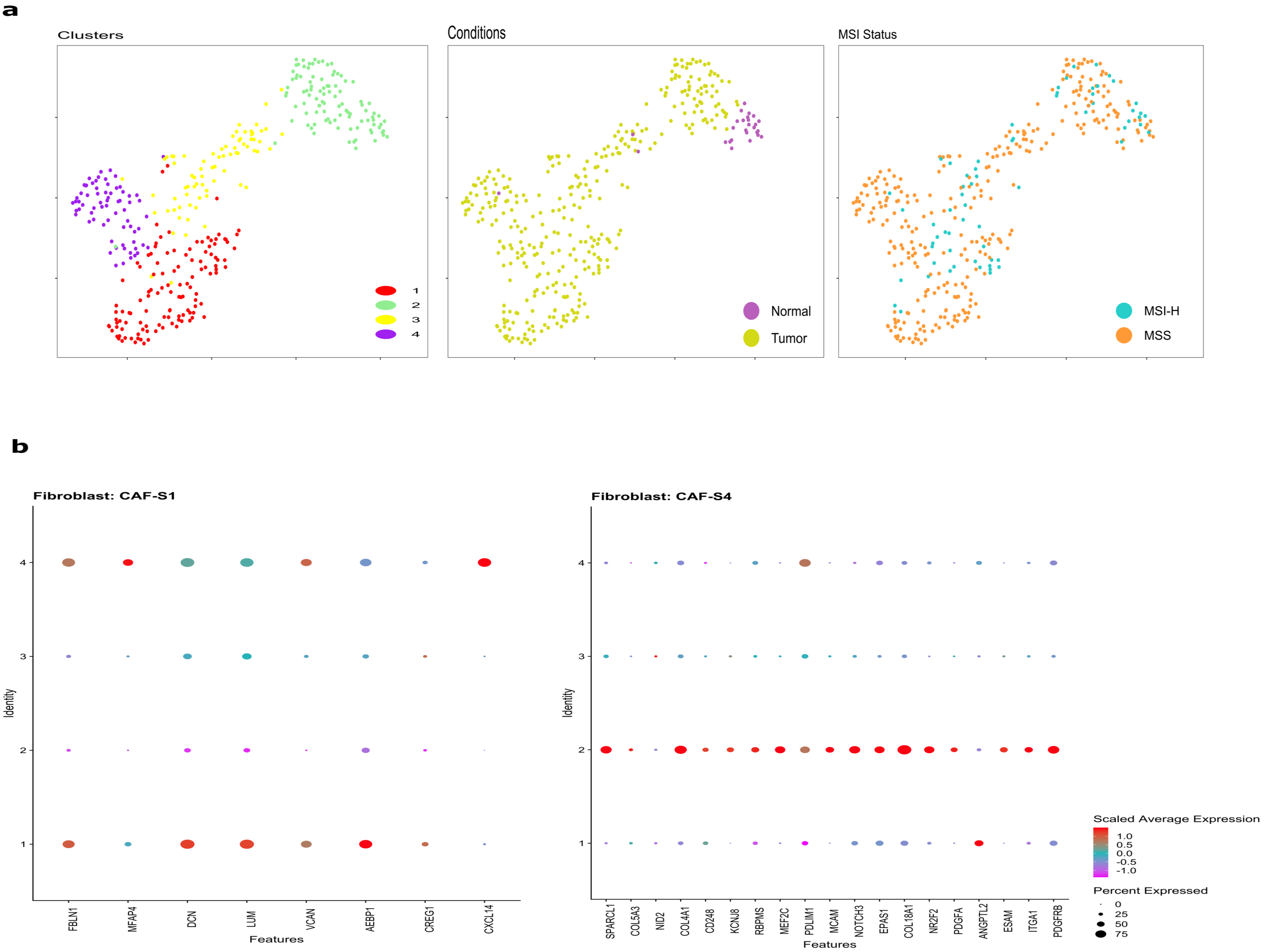
Supplementary figure 2. **a,** Heatmap of marker gene expression for the epithelial and tumor cells. **b,** UMAP visualization of computational analysis of differentiation status using CytoTRACE (see methods). **c,** Heatmap representation of Hallmark pathway analysis of epithelial phenotypes within 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

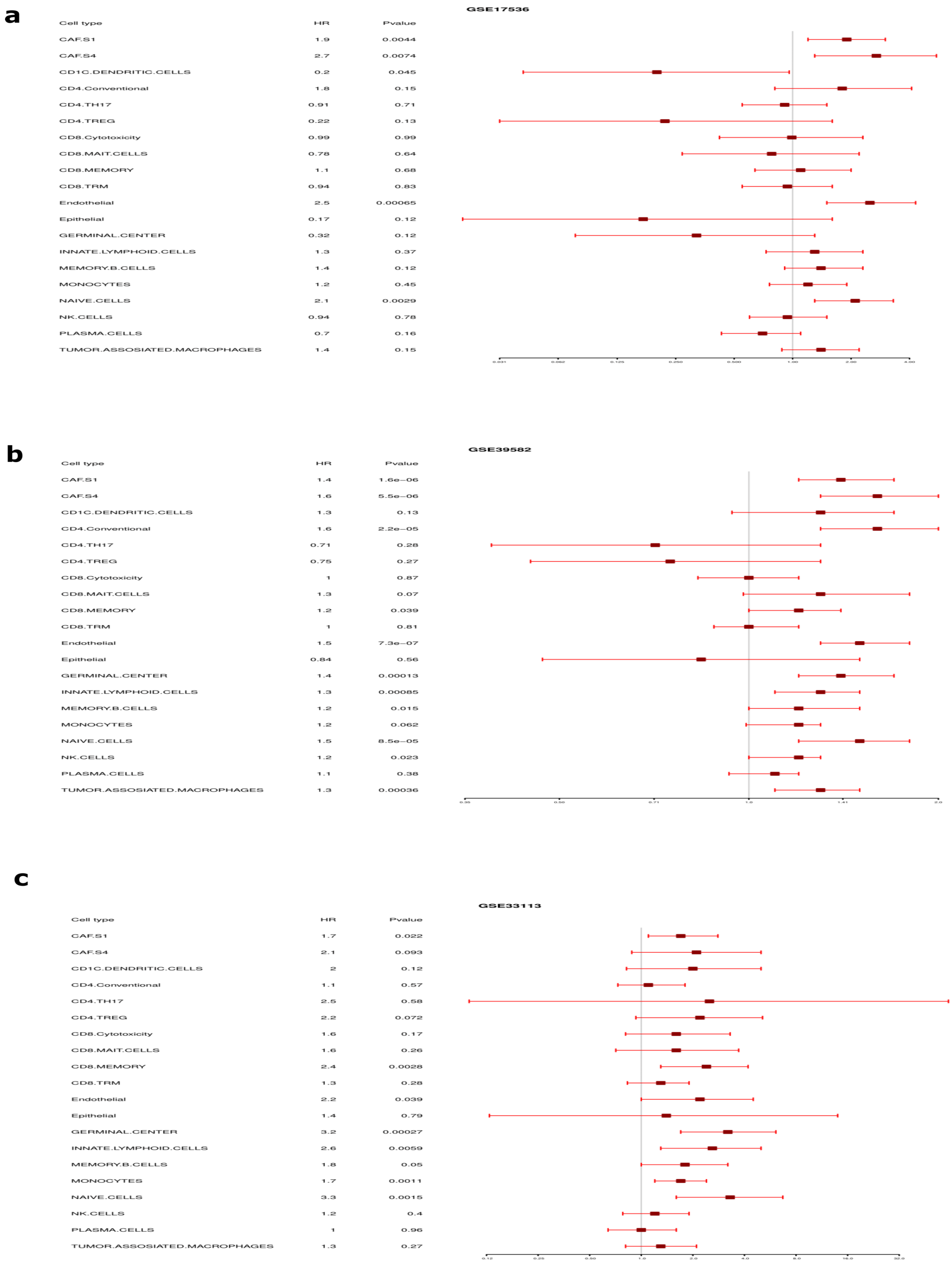
Supplementary figure 3. Pathway analysis (GO ontology) of gene-expressions specific to each cell-state in trajectory analysis stratified by malignant and non-malignant sample of origin.


a
Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
Condition
Normal
Tumor
MSI-Status
MSI-H
MSS
b
State
1
2
3
4
5
6
7
Component 1
Component 2
States
c
CMS1
CMS2
CMS3
CMS4
NA
Component 1
Component 2
Sc-CMS
d
nearestCMS.RF
CMS1
CMS2
CMS3
CMS4
Component 1
Component 2
Bulk CMS
e
Normal
Tumor
Component 1
Component 2
Condition
f
left
right
Component 1
Component 2
Location

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 MSI status. **b-f,** Differentiation trajectories of epithelial cells colored by differentiation state, CMS status, tumor status and colonic location.


a
Clusters
1
2
3
4
Conditions
Normal
Tumor
MSI Status
MSI-H
MSS
b
Fibroblast: CAF-S1
Identity
4
3
2
1
FBLN1
MFAP4
DCN
LUM
VCAN
AEBP1
CREG1
CXCL14
Features
Fibroblast: CAF-S4
Identity
4
3
2
1
SPARCL1
COL5A3
NID2
COL4A1
CD248
KCNJ8
RBPMS
MEF2C
PDLIM1
MCAM
NOTCH3
EPAS1
COL18A1
NR2F2
PDGFA
ANGPTL2
ESAM
ITGA1
PDGFRB
Features
Scaled Average Expression
1.0
0.5
0.0
-0.5
-1.0
Percent Expressed
0
25
50
75

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

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

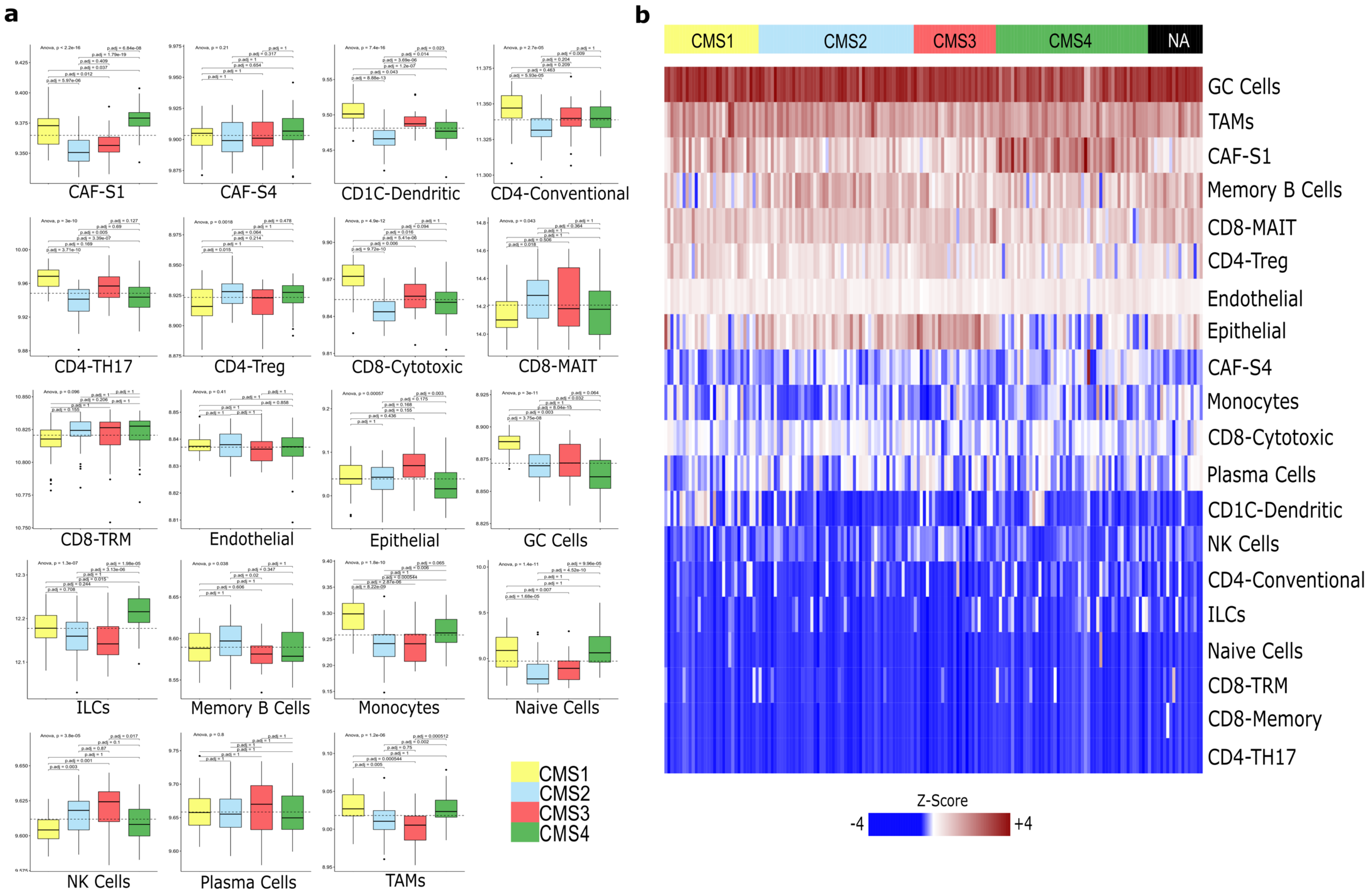
b

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

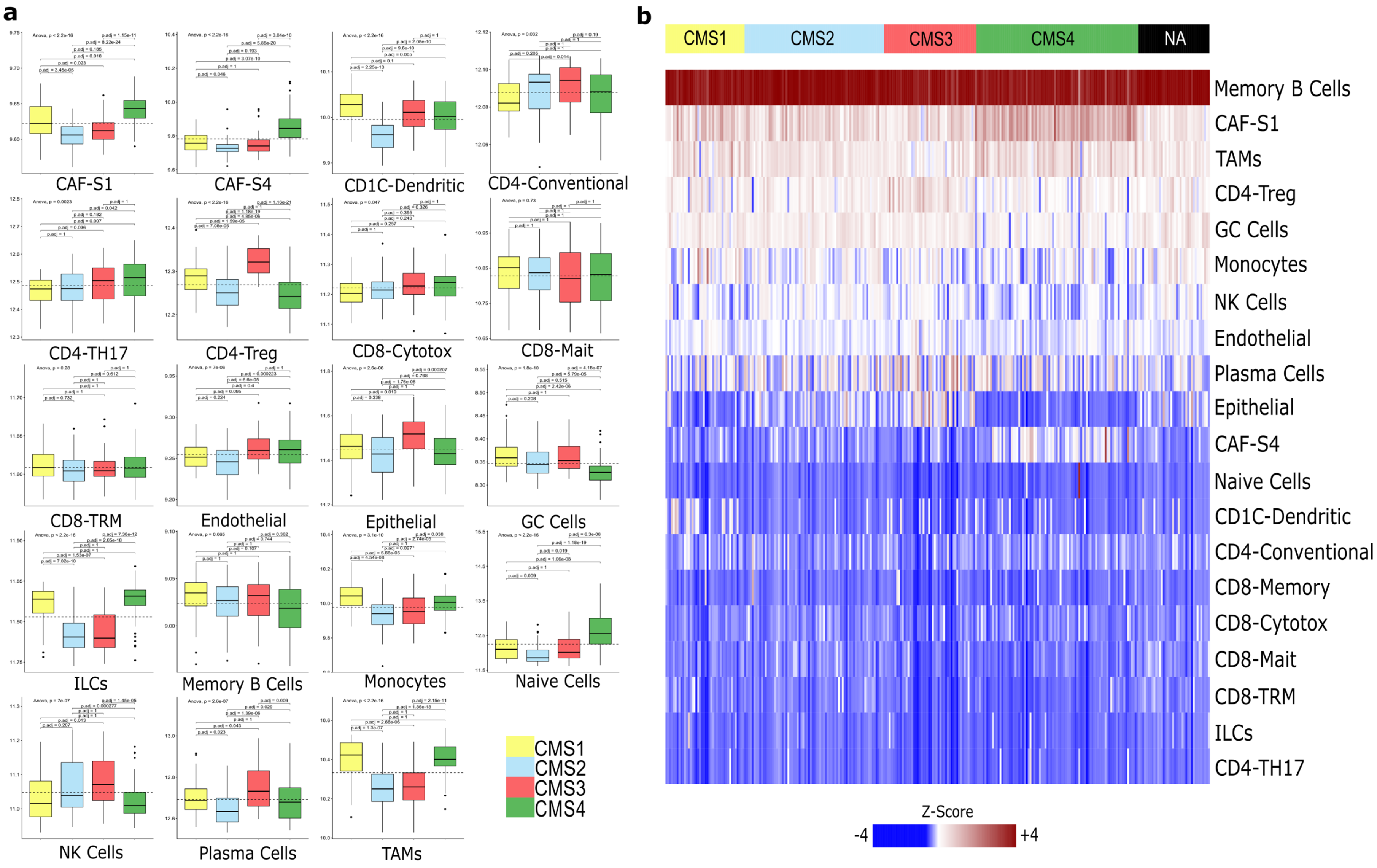
c

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

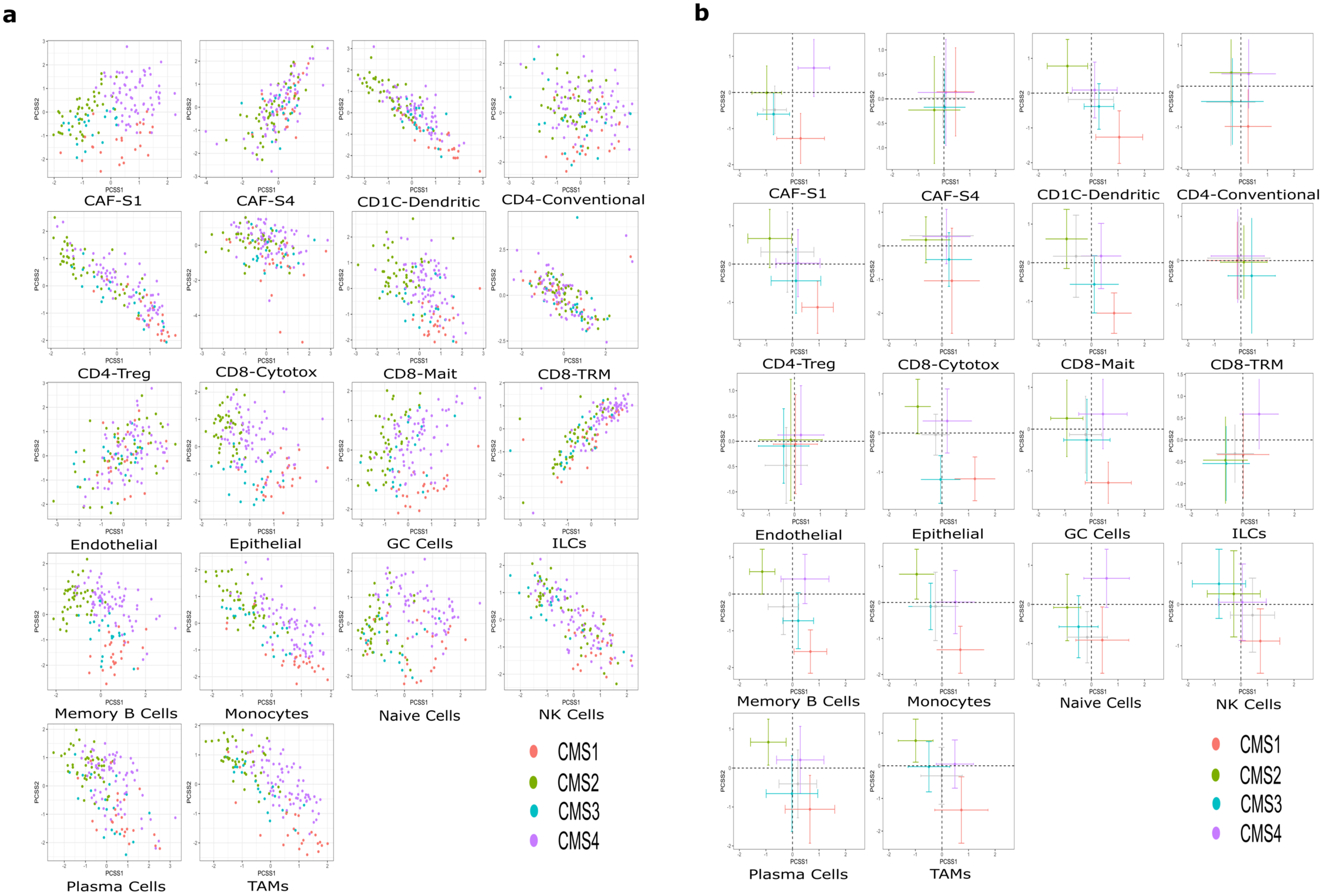
Supplementary Figure 6 Association between relative cell abundance and patient survival from microarray-based datasets. a. GSE17536 (n=177). b GSE39582 (n=585). c. GSE33113 (n=96). Note that CAF-S4 is not significant in GSE33113 (p=0.093).


a
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-TH17
CD4-Treg
CD8-Cytotoxic
CD8-MAIT
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
🟨 CMS1
🟦 CMS2
🟥 CMS3
🟩 CMS4
b
CMS1
CMS2
CMS3
CMS4
NA
GC Cells
TAMs
CAF-S1
Memory B Cells
CD8-MAIT
CD4-Treg
Endothelial
Epithelial
CAF-S4
Monocytes
CD8-Cytotoxic
Plasma Cells
CD1C-Dendritic
NK Cells
CD4-Conventional
ILCs
Naive Cells
CD8-TRM
CD8-Memory
CD4-TH17
Z-Score
-4
+4

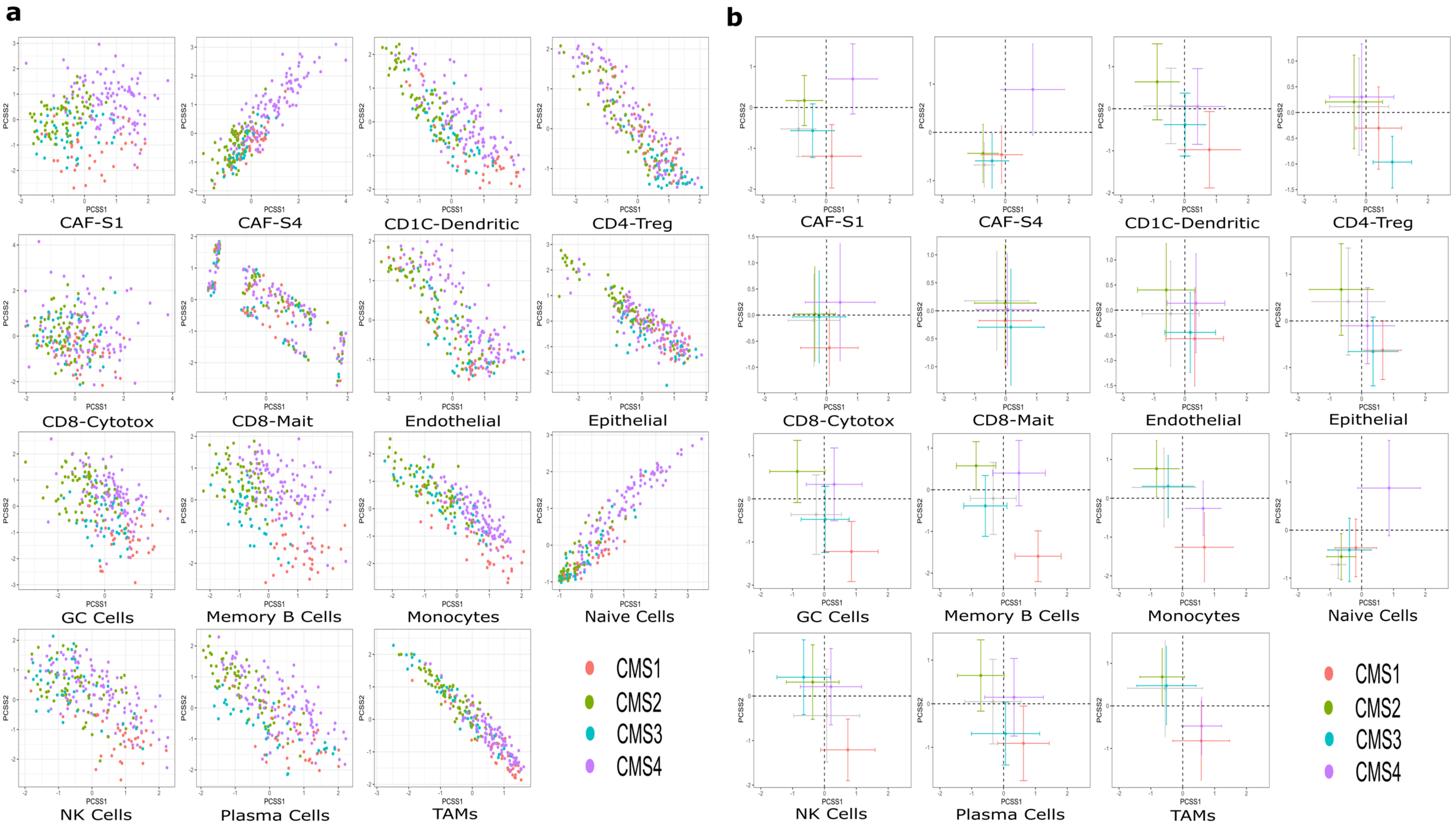
Supplementary Figure 7. Cell abundance prediction for each sample and projected on bulk CMS on GSE17536 (n= 177). a, Boxplots show the distribution of cell types within tumors with varying CMS status. The whiskers depict the 1.5 x IQR. The p-values for pairwise t-tests comparisons (with Benjamani-Hochberg correction) and ANOVA tests of cell abundance across CMS are shown in the figure. **b,** Deconvolution heatmap of different cell types by average expression using CIBERSORTx demonstrating cell type distribution within each CMS category. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.


a
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-TH17
CD4-Treg
CD8-Cytotox
CD8-Mait
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
CMS1
CMS2
CMS3
CMS4
b
CMS1
CMS2
CMS3
CMS4
NA
Memory B Cells
CAF-S1
TAMs
CD4-Treg
GC Cells
Monocytes
NK Cells
Endothelial
Plasma Cells
Epithelial
CAF-S4
Naive Cells
CD1C-Dendritic
CD4-Conventional
CD8-Memory
CD8-Cytotox
CD8-Mait
CD8-TRM
ILCs
CD4-TH17
Z-Score
-4
+4

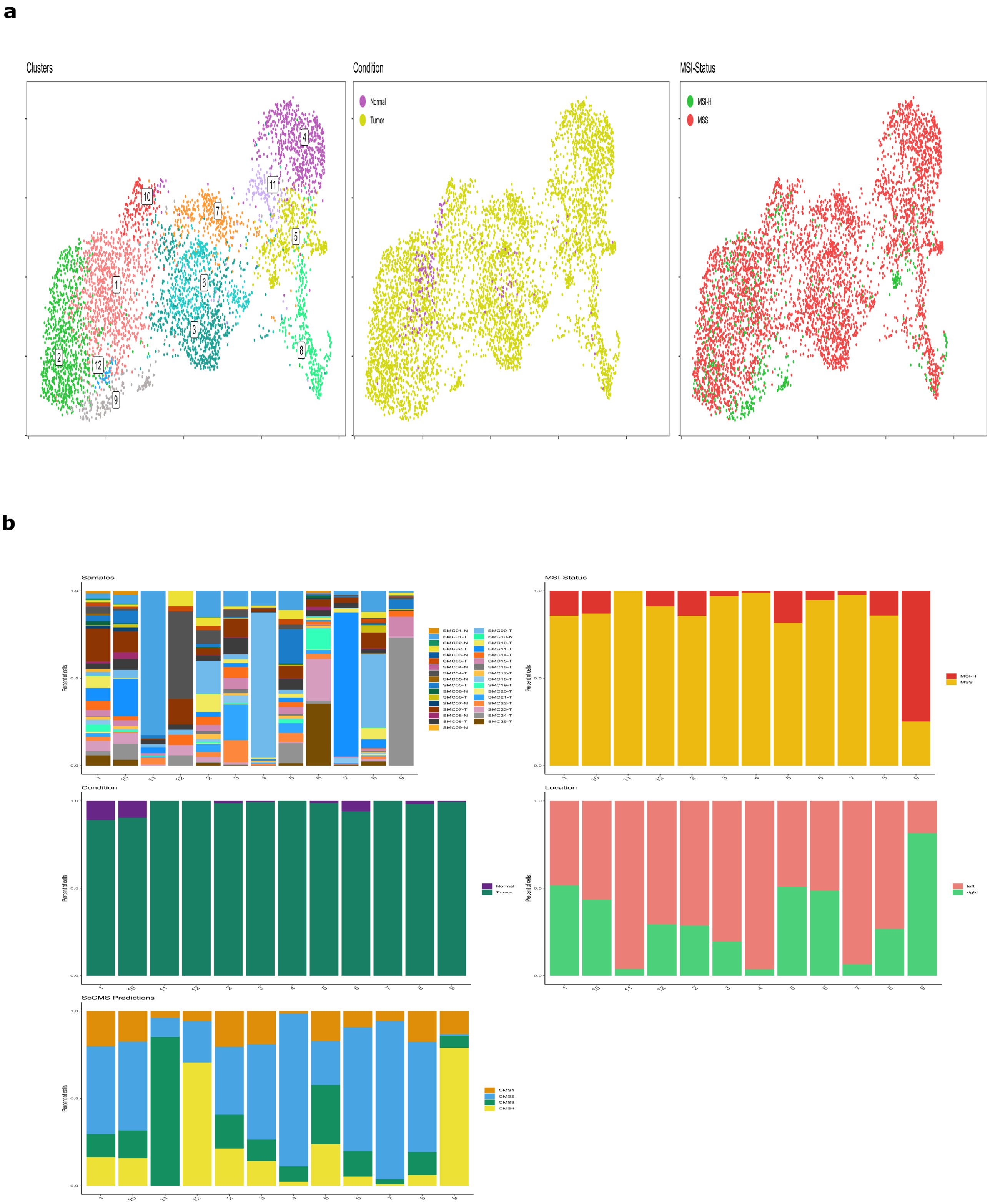
Supplementary Figure 8. Cell abundance prediction for each sample and projected on bulk CMS on GSE14333 (n= 290). a, Boxplots show the distribution of cell types within tumors with varying CMS status. The whiskers depict the 1.5 x IQR. The p-values for pairwise t-tests comparisons (with Benjamani-Hochberg correction) and ANOVA tests of cell abundance across CMS are shown in the figure. **b,** Deconvolution heatmap of different cell types by average expression using CIBERSORTx demonstrating cell type distribution within each CMS category. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.


a
b
PCSS1
PCSS2
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Conventional
CD4-Treg
CD8-Cytotox
CD8-Mait
CD8-TRM
Endothelial
Epithelial
GC Cells
ILCs
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
● CMS1
● CMS2
● CMS3
● CMS4

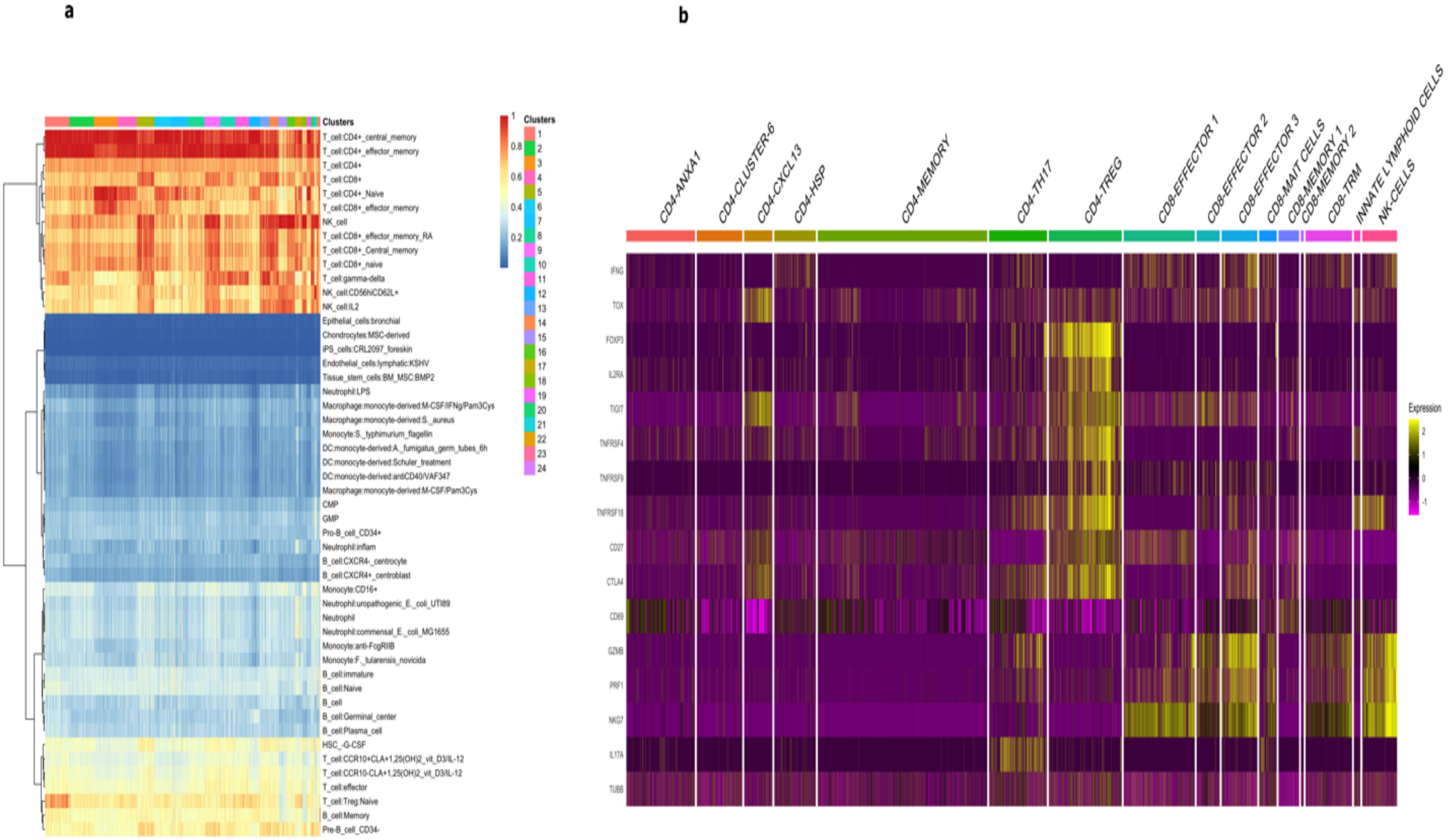
Supplementary Figure 9. Continuous scores for CRC dataset GSE17536 (n= 177). a, Principal component analysis plot showing PCSS1 and PCSS2 continuous scores reported by CMS classification across cell types show minimal separation in the top 2 principal components. **b,** All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Note that the cell types largely form a continuum along CMS status and are not clustered in discrete quadrants separate from one another. Cells and markers are colored by bulk CMS status accordingly to the tumor sample of origin. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.


a
b
CAF-S1
CAF-S4
CD1C-Dendritic
CD4-Treg
CD8-Cytotox
CD8-Mait
Endothelial
Epithelial
GC Cells
Memory B Cells
Monocytes
Naive Cells
NK Cells
Plasma Cells
TAMs
PCSS1
PCSS2
● CMS1
● CMS2
● CMS3
● CMS4

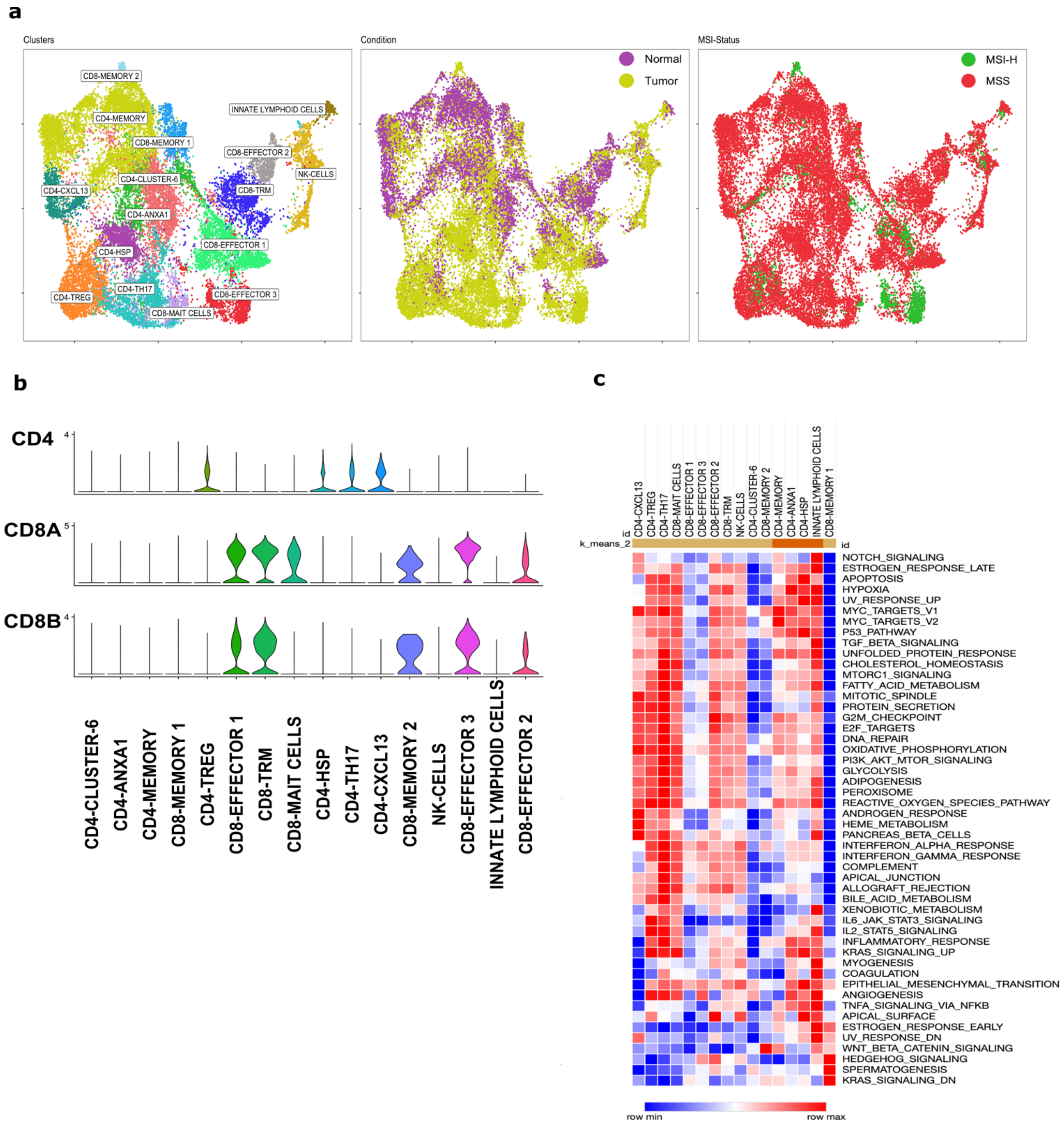
Supplementary Figure 10. Continuous scores for CRC dataset GSE14333 (n= 290). a, Principal component analysis plot showing PCSS1 and PCSS2 continuous scores reported by CMS classification across cell types show minimal separation in the top 2 principal components. **b,** All cell types projected on four quadrants representing CMS1-4 using PCSS1 and PCSS2 scores. Note that the cell types largely form a continuum along CMS status and are not clustered in discrete quadrants separate from one another. Cells and markers are colored by bulk CMS status accordingly to the tumor sample of origin. ILCs = Innate lymphoid cells, GC Cells = Germinal Center B Cells, NK Cells = Natural Killer Cells, TAMs = Tumor Associated Macrophages.



Supplementary figure 9. T/NK cell data from Lee et al. a, Reclustering of T/NK cells and coloring by clusters, tumor and normal status, MSI status and samples. **b**, Bar chart representation of cells colorized by our samples of origin, MSI status, tumor status, colonic location, CMS status.

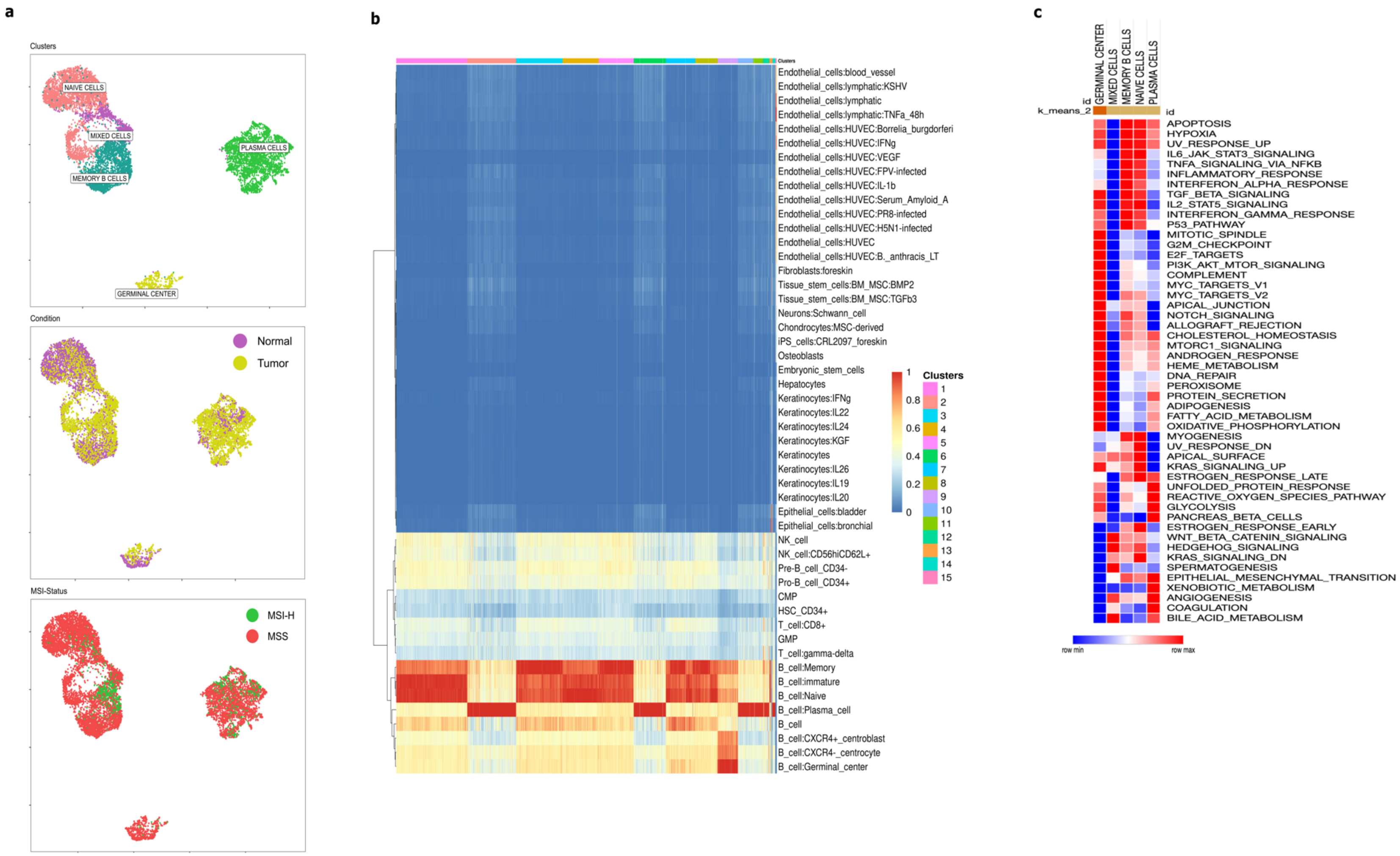


Supplementary figure 10. T cell expression analyses across T/NK cells . a, SingleR heatmap cell type identification within each cluster. Note that genes associated with T cells show increased expressions, confirming the quality of the data. Note doublets were removed from the further analysis. **b,** T/NK cells gene specific expression to identify T cell heterogeneity using published literature (see methods).



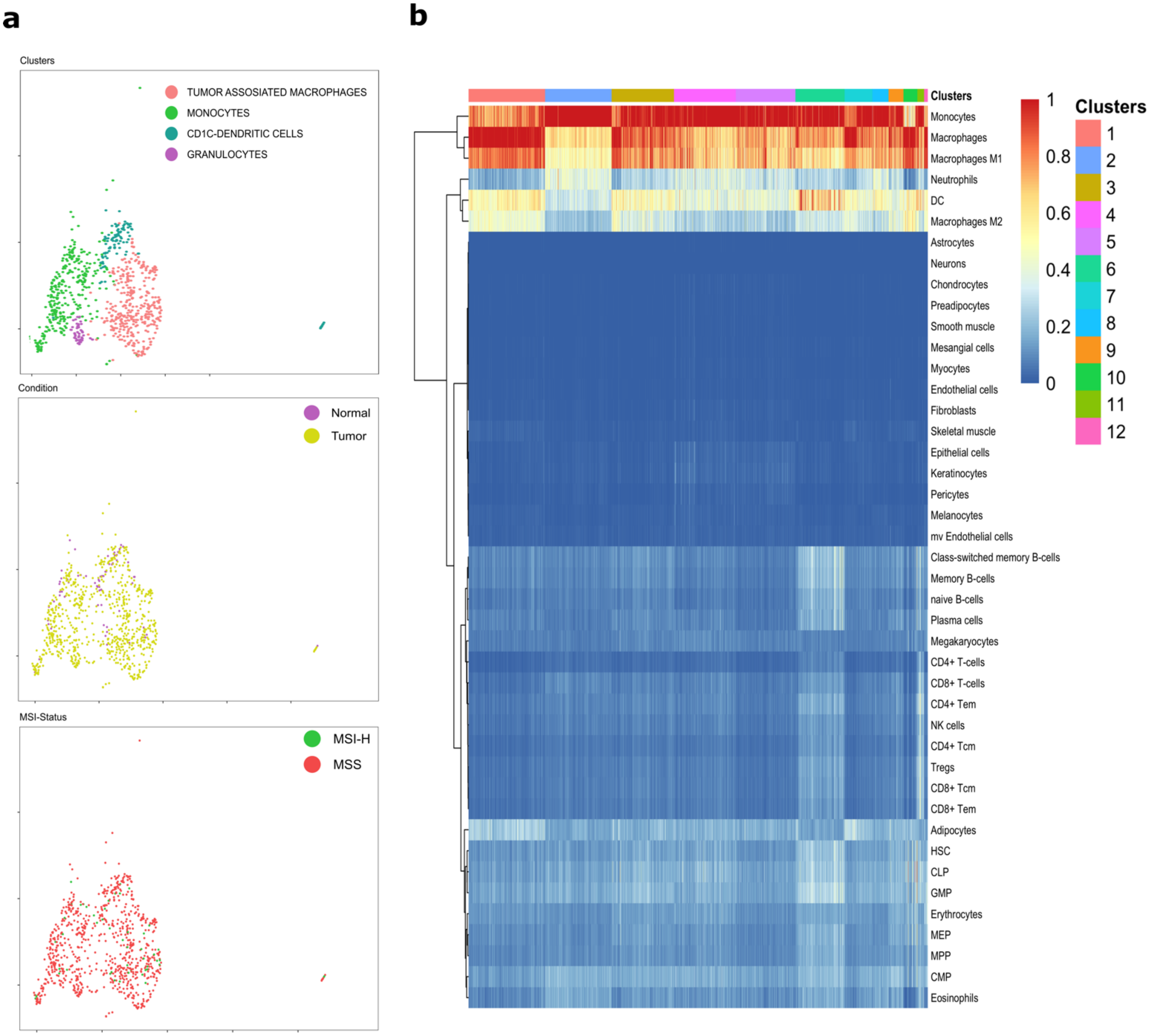
Extended Data Figure 1. **T cell identification and characterization. a**, UMAP reclustering of T cells colored by cell phenotype, tissue malignancy status, and sample of origin. **b**, Violin plots showing the differential expression of T cell-specific marker genes between CD4 and CD8 phenotypes. **c**, 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.


a
Clusters
NAIVE CELLS
MIXED CELLS
MEMORY B CELLS
PLASMA CELLS
GERMINAL CENTER
Condition
Normal
Tumor
MSI-Status
MSI-H
MSS
b
Clusters
Endothelial_cells:blood_vessel
Endothelial_cells:lymphatic:KSHV
Endothelial_cells:lymphatic
Endothelial_cells:lymphatic:TNFa_48h
Endothelial_cells:HUVEC:Borrelia_burgdorferi
Endothelial_cells:HUVEC:IFNg
Endothelial_cells:HUVEC:VEGF
Endothelial_cells:HUVEC:FPV-infected
Endothelial_cells:HUVEC:IL-1b
Endothelial_cells:HUVEC:Serum_Amyloid_A
Endothelial_cells:HUVEC:PR8-infected
Endothelial_cells:HUVEC:H5N1-infected
Endothelial_cells:HUVEC
Endothelial_cells:HUVEC:B._anthracis_LT
Fibroblasts:foreskin
Tissue_stem_cells:BM_MSC:BMP2
Tissue_stem_cells:BM_MSC:TGFb3
Neurons:Schwann_cell
Chondrocytes:MSC-derived
iPS_cells:CRL2097_foreskin
Osteoblasts
Embryonic_stem_cells
Hepatocytes
Keratinocytes:IFNg
Keratinocytes:IL22
Keratinocytes:IL24
Keratinocytes:KGF
Keratinocytes
Keratinocytes:IL26
Keratinocytes:IL19
Keratinocytes:IL20
Epithelial_cells:bladder
Epithelial_cells:bronchial
NK_cell
NK_cell:CD56hiCD62L+
Pre-B_cell_CD34-
Pro-B_cell_CD34+
CMP
HSC_CD34+
T_cell:CD8+
GMP
T_cell:gamma-delta
B_cell:Memory
B_cell:immature
B_cell:Naive
B_cell:Plasma_cell
B_cell
B_cell:CXCR4+_centroblast
B_cell:CXCR4-_centrocyte
B_cell:Germinal_center
1
0.8
0.6
0.4
0.2
0
Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
c
GERMINAL CENTER
MIXED CELLS
MEMORY B CELLS
NAIVE CELLS
PLASMA CELLS
id
k_means_2
id
APOPTOSIS
HYPOXIA
UV_RESPONSE_UP
IL6_JAK_STAT3_SIGNALING
TNFA_SIGNALING_VIA_NFKB
INFLAMMATORY_RESPONSE
INTERFERON_ALPHA_RESPONSE
TGF_BETA_SIGNALING
IL2_STAT5_SIGNALING
INTERFERON_GAMMA_RESPONSE
P53_PATHWAY
MITOTIC_SPINDLE
G2M_CHECKPOINT
E2F_TARGETS
PI3K_AKT_MTOR_SIGNALING
COMPLEMENT
MYC_TARGETS_V1
MYC_TARGETS_V2
APICAL_JUNCTION
NOTCH_SIGNALING
ALLOGRAFT_REJECTION
CHOLESTEROL_HOMEOSTASIS
MTORC1_SIGNALING
ANDROGEN_RESPONSE
HEME_METABOLISM
DNA_REPAIR
PEROXISOME
PROTEIN_SECRETION
ADIPOGENESIS
FATTY_ACID_METABOLISM
OXIDATIVE_PHOSPHORYLATION
MYOGENESIS
UV_RESPONSE_DN
APICAL_SURFACE
KRAS_SIGNALING_UP
ESTROGEN_RESPONSE_LATE
UNFOLDED_PROTEIN_RESPONSE
REACTIVE_OXYGEN_SPECIES_PATHWAY
GLYCOLYSIS
PANCREAS_BETA_CELLS
ESTROGEN_RESPONSE_EARLY
WNT_BETA_CATENIN_SIGNALING
HEDGEHOG_SIGNALING
KRAS_SIGNALING_DN
SPERMATOGENESIS
EPITHELIAL_MESENCHYMAL_TRANSITION
XENOBIOTIC_METABOLISM
ANGIOGENESIS
COAGULATION
BILE_ACID_METABOLISM
row min
row max

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**, SingleR heatmap demonstration of B cell distribution within each cluster. **c**, 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 nonmalignant tissue.



Extended Data Figure 3. **Reclustering of the myeloid cell compartment. a**, UMAP depiction of myeloid cell reclustering colored by subtype, malignancy status, and sample of origin. **b**, SingleR heatmap demonstration of myeloid cell distribution within each cluster. scCMS, single-cell consensus molecular subtyping; bulk CMS, consensus molecular subtyping on Bulk RNA-seq data.

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