Dissecting heterogeneity of tumor microenvironment in colorectal cancer using high-resolution single-cell atlas

Valentin Marteau[1, 11], Niloofar Nemati[1, 11], Kristina Handler[2, 11], Deeksha Raju[2], Erika Kvalem Soto[1], Georgios Fotakis[1], Sandro Carollo[1], Nina Böck[1], Alexander Kirchmair[1], Alexandra Scheiber[3], Arno Ammann[3], Andreas Seeber[3], Elisabeth Gasser[4], Lena Horvath[3], Hassan Fazilaty[5], Tomas Valenta[5, 8], Gregor Sturm[1,7], Dietmar Rieder[1], Sieghart Sopper[3], Andreas Pircher[3], Paul Ziegler[9], Markus Tschurtschenthaler[10], Dominik Wolf [3], Isabelle Arnold[2, 6, 12*], Stefan Salcher[3, 12*], Zlatko Trajanoski[1, 12*]

[1]Biocenter, Institute of Bioinformatics, Medical University of Innsbruck, Austria
[2]Institute of Experimental Immunology, University of Zurich, Switzerland
[3]Department of Internal Medicine V, Haematology & Oncology, Comprehensive Cancer Center Innsbruck (CCCI) and Tyrolean Cancer Research Institute (TKFI), Medical University of Innsbruck (MUI), Innsbruck, Austria
[4]Department of Visceral, Transplant and Thoracic Surgery, Medical University Innsbruck, Anichstrasse 35, 6020, Innsbruck, Austria
[5]Department of Molecular Life Sciences, University of Zürich, Zürich, Switzerland
[6]Comprehensive Cancer Center Zürich, Zürich, Switzerland
[7]Boehringer Ingelheim International Pharma GmbH & Co KG, Biberach, Germany
[8]Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[9]Senckenberg. Institute of Pathology, Goethe University Frankfurt, Frankfurt, Germany
[10]Translational Cancer Research and Institute of Experimental Cancer Therapy, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany
[11]These authors contributed equally
[12]These authors contributed equally
*Correspondence: stefan.salcher@i-med.ac.at
*Correspondence: zlatko.trajanoski@i-med.ac.at

SUMMARY

Colorectal cancer (CRC) features molecular heterogeneity and differing immune cell infiltration patterns, affecting how patients respond to therapy. However, a comprehensive understanding of immune cell phenotypes across patient subgroups is missing. Here, we dissect the CRC tumor microenvironment by integrating 4,27 million single cells from 1670 samples and 650 patients across 48 studies (76 datasets) representing 7 billion expression values. These samples encompass the full spectrum of disease progression, from normal colon to polyps, primary tumors, and metastases, covering both early and advanced stages of CRC. We present a high-resolution view of CRC with 62 major cell types or states, revealing different cell-type composition patterns in CRC subtypes. The atlas allows for refined tumor classification and patient stratification into four immune phenotypes: immune desert, myeloid, B cell, and T cell subtypes. Using bulk samples with genomic and clinical information, we identify cellular components associated with tumor histology and genotypes. These findings could have important implications for developing improved cancer immunotherapy approaches in CRC.

KEYWORDS

Single-cell sequencing, cell-cell communication, patient stratification, therapy response

INTRODUCTION

Colorectal cancer (CRC) remains the most common gastrointestinal cancer and the second leading cause of cancer-related death according to the GLOBOCAN data[1]. The growing incidence in some countries reflects a modification in lifestyle and its consequences related with "Westernization" such as diet, obesity, physical inactivity, and alcohol consumption. Similarly, "Westernization" of diets and lifestyle have been attributed to the alarming growing incidence of early-onset CRCs (individuals <50 years of age)[2]. Despite the improvement of early detection screening policies and the development of many novel treatments, the median overall survival (OS) for patients with metastatic CRC (mCRC) is ~30 months[3]. Progress in targeted therapy for mCRC has been limited, relative to other solid cancers like lung cancer or melanoma[4]. The genetic heterogeneity[5] as well as paucity of drugs targeting common driver genes, pose considerable challenges for developing precision oncology approaches for patients with CRC. Moreover, CRC seems to be refractory to therapy with immune checkpoint inhibitors (ICI) with the notable exception of 4% of patients with tumors characterized by microsatellite instability (MSI) or POLE proofreading mutations[6]. This is somehow paradoxical since CRCs are known to be under immunological control, as we have shown in the past[7]. Thus, there is an urgent need to explore new immunomodulatory strategies for the vast majority of CRC patients with microsatellite stable (MSS) tumors. Strategies that convert the "immune cold" tumor microenvironment (TME) to "immune hot" could sensitize cancers to ICI therapy. A number of clinical trials with drug combinations have been carried out to overcome resistance to immunotherapy in MSS CRC[8,9]. The benefits of using a dual combination of ICIs (CTLA-4 and PD-1 blockade) or adding ICIs to chemotherapy appear to be very limited[8,10]. Combination strategies that exploit synergistic potential of ICI and targeted drugs showed increase of the anti-tumor activity, however, the objective response rates were disappointingly low[8]. Thus, to address this unmet clinical need a better understanding of the immunosuppressive TME and its heterogeneity is of utmost importance.

The intrinsic complexity of the interaction of the two interwoven systems, the tumor and the immune system, poses considerable challenges and requires comprehensive approaches to interrogate cancer immunity during tumor initiation and progression, and following therapeutic modulation thereof. A number of established and novel high-throughput technologies enable the generation of the necessary data and thereby provide the basis for mechanistic understanding. Specifically, next-generation sequencing (NGS) of cancer samples provide large data sets that can be mined for immunologically relevant parameters (see our reviews[11,12]). Using these tools we were able to chart the TME in CRC[13] and developed The Cancer Immunome Atlas[14]. While these attempts uncovered the enormous heterogeneity of MSS CRC and revealed unique cancer cell and TME subtypes with prognostic and therapeutic relevance, the advent of single-cell RNA sequencing (scRNA-seq) technologies opened fundamental new avenues to dissect cancer heterogeneity. We recently developed a novel computational

80 approach to dissect the cellular diversity in the TME at high resolution by integrating expression profiles
81 of single cells from multiple studies using cutting-edge artificial intelligence (AI) methods[15].
82
83 Here we used our approach to build a high-resolution single-cell transcriptome atlas for CRC with 4.27
84 million cells from 650 patients and 1,670 samples. This high-resolution atlas offers superior coverage
85 of histological and clinical variables and provides a unique resource for dissecting the cellular diversity
86 in the TME. We then used bulk RNA-seq data from The Cancer Genome Atlas (TCGA)[5] to evaluate the
87 association of atlas-derived single-cell derived transcriptomic signatures with survival data. The
88 exploitation of this high-resolution single-cell transcriptome atlas for CRC will provide novel biological
89 insights that will be the basis to improve therapy in MSS CRC patients.
90

RESULTS

Generation of a core large-scale CRC single-cell atlas

We developed a core CRC atlas by compiling scRNA-seq data from 48 studies and 76 datasets comprising 4.27 million cells from 650 patients and 1,670 samples (Figure 1A). This comprehensive CRC single-cell atlas integrates expert-curated, quality-assured and pre-analyzed transcriptomic data from publicly available studies encompassing the full spectrum of disease progression, from normal colon to polyps, primary tumors, and metastases, covering both early and advanced stages of CRC (see Methods). Important study characteristics are summarized in Table S1. In total, the core atlas includes transcriptomic data from 487 patients with CRC, 99 patients with polyps, and 64 control individuals. Of the 487 CRC patients, 254 were histopathologically annotated as colon (COAD), 63 as rectal cancer (READ), and 170 were not otherwise specified. CRC samples include tissue of the primary tumor (n=674) or metastasis (n=175) that were obtained by surgical resection. Among the control samples, 284 were derived from distant non-malignant tissue of colon tumor patients, of which 220 have a patient-matched tumor sample. 186 non-tumor control samples as well as 116 polyp and 90 matched normal samples were obtained through patients undergoing colonoscopy. Further, 103 blood samples of CRC patients as well as 24 samples derived from tumor affected lymph nodes, with another 18 non-tumor affected lymph node samples were included.

We clustered the CRC patients' disease stages as early (UICC stage I – II) *versus* advanced (UICC III-IV) disease, as not all studies provided sufficient information on tumor stages. Of the control patients none had a history of IBD. Overall, the core atlas integrates 4,264,946 single cells, which are annotated to 11 coarse cell type identities and 62 major cell subtypes or cell states (e.g. dividing cells) based on previously established canonical single-cell signatures including 1,146,196 epithelial cells, 1,552,108 T/NK cells, 804,042 B/Plasma cells, 385,589 myeloid cells, and 33,9767 stromal and endothelial cells (Figure 1B).

Cancer cells (in total 600,740 cells, from primary tumor and metastatic tissue) in the atlas showed high heterogeneity of their transcriptomic profiles (Figure 1B). Due to the large patient cohort, we were able to apply high-resolution cancer cell classification based on their specific marker gene expression signatures (Figure S1D). We divided the following main clusters Colonocyte-like (*FABP1, CEACAM7, ANPEP, SI*), BEST4 (BEST4*, OTOP2, SPIB*), Goblet-like (*MUC2, FCGBP, ZG16, ATOH1*), Crypt-like (*LGR5, SMOC2*), and TA-like (*TOP2A, UBE2C, PCLAF, HELLS, TK1*) (Figure 1B, Figure S1D). There are highly mitotic/proliferative clusters (*LGR5, TOP2A, MKI67*) of both COAD and READ that may resemble highly aggressive and invasive cancer cells. In addition, *EPCAM* expressing cells in the blood of CRC patients were annotated as circulating cancer cells.

128 **Single-cell composition of the TME reveals distinct CRC tumor immunotypes**
129 Next, we stratified CRC patients based on infiltration patterns of their TME using the high-resolution
130 atlas. Unsupervised clustering on batch-corrected cell-type fractions revealed four distinct tumor
131 immune phenotypes (Figure 2A): 1) immune deserted tumors (ID; i.e. no significant immune cell
132 infiltration but high cancer cell fraction); 2) the subtype of tumors with B cell dominance (B; B cell,
133 plasma IgA and IgG cell); 3) the subtype of tumors with myeloid dominance (M;
134 macrophage/monocyte); and 4) subtype of tumors with T cell dominance (T; $CD8^+$, $CD4^+$, T regulatory
135 cells). Using logistic regression, we did not find any association of the different patient strata to tumor
136 stages (early *vs.* advanced) or sex.
137
138 We identified gene signatures based on the patient immune infiltration groups by selecting the top 15
139 differentially expressed genes. These gene signatures were then used to conduct a survival analysis
140 utilizing data from the TCGA COAD-READ cohorts (n = 629). The Kaplan-Meier curves in Figure 2C
141 demonstrate that higher expression of the B (p-value = 0.0015), M (p-value = 0.0096), and T (p-value =
142 0.00078) signatures is significantly correlated with improved overall survival. The analysis showed that
143 higher expression of B, M, and T gene signatures is significantly linked to better overall survival.
144
145 **Analysis of cell-cell communication uncovers interactions between different cell types in the tumor**
146 **microenvironment (TME)**
147 Differences in cellular crosstalk were analyzed using LIANA[29] framework, which leverages a consensus
148 of expert-curated ligand-receptor databases, as detailed in the methods section. In this study, our primary
149 focus is on the interactions between cancer cells and the main immune cell types. Specifically, we
150 examined the top 10 differentially expressed ligands in cancer cells (Figure 3). *CD69,* which is known
151 to be involved in the early activation and proliferation of T lymphocytes[31], was found to be upregulated
152 in CD8+ T cells, regulatory T cells, and CD4+ T cells. We observed increased signaling of several
153 chemokine receptors. This finding may indicate that cancer cell-secreted chemokine gradients may be
154 contributing to immune cell infiltration into the tumor microenvironment.
155
156 **Integration of bulk RNA-seq data reveals genotype-immune phenotype associations**
157 scRNA-seq provides an unprecedented view on the cellular heterogeneity in the TME. However, the
158 majority of the scRNA-seq studies lack both comprehensive cancer genotype information and survival
159 data. In contrast, the TCGA reference dataset includes both of these, alongside bulk RNA-sequencing
160 data. Using the computational method SCISSOR[36], we integrated bulk RNA-seq data from TCGA with
161 single-cell transcriptomic signatures to assess the relationship between specific genotypes and survival
162 outcomes in CRC. We evaluated the association of cell type transcriptomic signatures derived from the
163 atlas with genotype and survival data from the TCGA reference dataset, which includes 629 patients
164 (UICC I-IV)[5]. Building on our previous work that identified genomic features influencing immune

165 phenotypes in a pan-cancer study using bulk RNA-seq data[37], we leveraged the high resolution of the
166 single-cell CRC atlas for a more detailed analysis of these determinants.
167
168 We next analyzed the cell type transcriptomic signatures and their association with survival in 629
169 patients from the TCGA cohort. Overall, the presence of high levels of Plasma IgA, $CD4^+$ T cells $CD8^+$
170 T cells and B cells showed a prominent association with improved survival in CRC patients. Conversely,
171 CRC patients with tumors enriched in Schwann cells were associated with worst survival outcomes, as
172 also reported previously[38]. Notably, high infiltration of various endothelial cells and fibroblasts also
173 emerged as strong negative predictors of survival (Figure 4A). Mutations in *BRAF, KRAS* were
174 associated with distinct immune infiltrates (Figure 4B-C). For example, cDC1 and cDC2 showed high
175 infiltration patterns in CRC patients with BRAF mutations and same for Plasma IgM where the patterns
176 were present in both BRAF and KRAS mutations. In contrast, mutations in BRAF were associated with
177 low Plasma IgA infiltration, whereas KRAS mutations correlated with high Plasma IgA infiltration. MSI
178 and MSS status were associated with opposite cell type patterns (Figure 4D-E).
179
180

181 **DISCUSSION**

182 We built a large-scale atlas of single-cell transcriptomes of CRC through integration of 4.27 million
183 cells from 650 patients and 1,670 samples representing 7 billion expression values. Reduction of dataset-
184 specific batch effects due to variation in experimental design and used platforms while retaining
185 biological information resulted in a high-quality reference atlas for CRC, offering superior coverage of
186 histological and clinical variables and thereby providing a unique resource for dissecting the cellular
187 diversity in the TME and generating hypotheses.

188

189 We provide a high-resolution view of the TME in CRC with 44 major cell types/states and show
190 different cell type composition patterns including more precise functional transcriptomic classification
191 of malignant epithelial cells. The single-cell composition of the TME in CRC enabled refined tumor
192 classification and patient stratification into four immune phenotypes: immune deserted, myeloid, B cell,
193 and T cell subtypes. These findings may have important implications for enabling cancer
194 immunotherapy in CRC. For example, combination therapies that target both, myeloid cells and
195 lymphoid cells could represent an immunotherapeutic strategy to treat M subtypes as shown recently in
196 a melanoma mouse model[41]. Similarly, given the heterogeneity of the intratumoral B cells and the
197 importance of tertiary lymphoid structures[42], further analysis of the B subtypes might open promising
198 therapeutic avenues by additional refined B cell targeting.

199

200 Furthermore, integration of bulk RNA-seq data from the TCGA CRC cohort uncovered cell subsets
201 associated with alterations in major driver genes, such as *BRAF*, *KRAS*, as well as with the CRC subtypes
202 MSI and MSS. Hence, we provide further evidence that genetic aberrations in cancer cells dictate the
203 immune contexture in tumors. This knowledge could be exploited to derive rationale for personalized
204 therapeutic combination strategies based on the underlying genetic profile of the tumor. The biological
205 insights we present here and future discoveries arising from the exploitation of the high-resolution CRC
206 atlas could provide the basis for developing combination therapies for CRC patients who are not
207 sufficiently responding to immune checkpoint blockers.

208

ACKNOWLEDGMENTS

This work was supported by the European Research Council (grant agreement No 786295 to ZT), by the Austrian Science Fund (FWF) (project I3978 to ZT). ZT is a member of the German Research Foundation (DFG) project TRR 241(INF). The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. VM was funded by 10.55776/DOC82. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

FIGURE LEGENDS

Figure 1. Schematic outline of the overall concept used in this study.

(A) Summary of the data integration and analysis workflow.

(B) Overview of the CRC atlas and the epithelial, immune, and stromal/endothelial components depicted as UMAP plots.

(C) Fractions of depicted cell types per scRNA-sequencing platform.

(D) Tissue samples and the corresponding cell types.

Figure 2. Tumor immune phenotypes in CRC.

(A) Patient characteristics and stratification of the tumor immune phenotypes. *Tumor type (histopathological)* refers to the histological subtypes as provided by the original datasets based on pathological assessment; *tumor type (transcriptomic)* is based on the most abundant transcriptomically annotated cancer cell-subtype in the scRNA-seq atlas.

(B) Differential activation of (B) PROGENy cancer pathways- (C) CytoSig cytokine signaling signatures in cancer cells between the four tumor immune phenotypes. Heatmap colors indicate the deviation from the overall mean, independent of tumor histology and stage. White dots indicate significant interactions at different false-discovery-rate (FDR) thresholds. Only cytokine signatures with an FDR < 0.1 in at least one patient group are shown.

See also Figure S2.

(C) Kaplan-Meier curves of the survival analysis according to the patient immune infiltration groups (B, desert, M, T), performed using the molecular data of the TCGA COAD-READ cohorts (n = 461).

Figure 3. Cellular crosstalk analysis

(A) Circos plot illustrating the interaction network between cancer cells and major immune cells in paired tumor and normal samples. The plot highlights the top 10 differentially expressed ligands from cancer cells. Red lines indicate interactions upregulated in tumors, while blue lines represent those upregulated in normal.

Figure 4. Association of cellular composition and distinct genotypes and survival in the TCGA data.

(A-E) *SCISSOR* analysis relating phenotypic information from bulk RNA-seq data from TCGA with single cells. UMAP plots indicate the position of cells positively (blue) or negatively (red) associated with mutation or better survival. A log2 ratio > 0 indicates a positive association with mutation, or better survival, respectively. Shown are cell-types with a log2 ratio significantly different from 0 at an FDR < 0.01 (paired Wilcoxon test).

(A) Association of cellular composition with overall survival.

(B) Association of cellular composition with BRAF mutation in CRC patients (n = 136).

(C) Association of cellular composition with KRAS mutation in CRC patients (n = 489).

(D) Association of cellular composition with MSI status in CRC patients (n = 177).

(E) Association of cellular composition with MSS status in CRC patients (n = 412).

262 **METHODS**

263 ***Generation of the atlas***

264 The data for the core atlas was from previously published studies[46-56], our own data, and 12 studies for
265 control purpose[50,57-62]. The selected studies were published between May 2017 and July 2024 and the
266 incorporated datasets were generated with six different sequencing platforms, including the most
267 commonly applied 10x Chromium (10x Genomics) as well as Smart-seq2[63], SMARTer (C1), scTrio-
268 seq2, GEXSCOPE (Singleron), TruDrop[64] and DNBelab C4[65]. We further integrated our own data
269 generated with the microwell-based BD Rhapsody scRNA-seq platform (BD Biosciences). We
270 specifically selected studies using comparable protocols for sample processing and data generation, such
271 as sequencing of whole cells. We did not exclude studies that applied flow cytometry-based cell-sorting
272 prior to sequencing, as these incorporate relevant information on rare cell types[48,49,66]. From non-CRC
273 studies we exclusively included those parts of the published data that were relevant for our atlas: from
274 the Garrido Trigo dataset [57] we only included colon samples, from Burclaff[60], Kong[58], Mazzurana[61],
275 Parikh[59], Smillie, Thomas, and Scheid[46] datasets we only used the control samples.

276

277 ***Integration of scRNA-seq datasets***

278 Individual datasets were merged into a single AnnData object. Since genome annotations partly differed
279 between the datasets, we mapped gene symbols to the ensembl ids of the used annotation version and
280 updated the symbols to the latest version of HGNC. If gene ids were missing from a dataset, the values
281 were filled with zeros. Genes that were missing in more than 19 datasets (25%) were excluded
282 altogether.

283 We integrated the datasets using the scANVI algorithm[68,69], as it has been demonstrated to be one of the
284 top-performing methods for atlas-level integration and to scale to >1M cells[70]. Since scANVI requires
285 cell-type annotations for at least one of the input datasets, we manually annotated two "seed" datasets
286 based on unsupervised clustering as described below. Raw counts were used as input for scANVI.
287 Counts from plate-based assays such as Smart-seq2 were scaled by the gene length as recommended on
288 the scvi-tools website. The scANVI model was initialized with a pre-trained scVI model[71], as
289 recommended in the scvi-tools tutorial. The scVI model was trained on the 6000 most highly variable
290 genes as determined with scanpy's[72] *pp.highly_variable_genes* with parameters *flavor="seurat_v3"*
291 and *batch_key="dataset"*. Each sample was considered as an individual batch for both scVI and
292 scANVI. Other than that the algorithms were run with default parameters.

293

294 ***Unsupervised clustering and cell-type annotation***

295 We computed UMAP embeddings[73] and unsupervised Leiden-clustering[74] with scanpy[72], based on a
296 cell-cell neighborhood graph derived from scANVI latent space. Coarse, lineage-specific clusters were

297 iteratively sub-clustered to identify cell-types at a more fine-grained resolution. Cell type clusters were
298 annotated based on previously reported marker genes[57,75,76]
299
300 ***Comparing cell-type abundances***
301 Since comparing cell-type fractions between groups is challenging due to different characteristics of the
302 datasets and the inherent compositional nature of cell-type fractions, we applied the scCODA [77] model,
303 which addresses this issue. We were interested in the differences between conditions (COAD *vs.*
304 READ). To this end, we ran the scCODA model with the formula ~ *condition* + *tumor_stage* + *dataset*
305 with 500,000 iterations using "cancer cells" as the reference cell-type, where *tumor_stage* is a binary
306 vector classifying datasets into early (stages I-II) and advanced (stages III-IV), and dataset is a
307 categorical vector encoding the different datasets. The final result shows *credible effects* with a false-
308 discovery-rate (FDR) of 0.1.
309
310 ***Patient stratification***
311 We stratified patients into immune phenotypes based on immune cell-type fractions. We selected all
312 patients with primary tumor samples and excluded enriched datasets. Neutrophil fractions were
313 excluded, since they are not appropriately captured in the majority of datasets. Cell-type fractions of
314 primary tumor samples were loaded into a patient × cell-type AnnData container. Dataset-specific batch-
315 effects were removed using a linear model as implemented in *scanpy.pp.regress_out*. Patients were
316 clustered using graph-based Leiden clustering with the "correlation" distance metric for computing the
317 neighborhood graph. Patient clusters were labeled according to their predominant cell-types. In addition
318 to the histological subtypes based on the annotation of the original datasets, we annotated tumor types
319 based on the transcriptomics data according to the most abundant cancer cell cluster.
320
321 ***Differential gene expression testing***
322 We used DESeq2[78] on pseudo-bulk samples for differential expression testing which has been
323 demonstrated to perform well and properly correct for false discoveries[79]. For each cell-type and patient,
324 we summed up transcript counts for each gene. Pseudo-bulk samples consisting of fewer than 10 cells
325 were discarded. We compared primary tumor samples from tumor *vs.* normal (*sample type*), including
326 the dataset as a covariate. For comparisons between multiple groups, we used contrasts with sum-to-
327 zero coding. P-values were adjusted for multiple hypothesis testing with independent hypothesis
328 weighting (IHW)[80].
329
330 ***Pathway, TF and cytokine signaling signatures***
331 We performed pathway, transcription factor (TF), and cytokine signaling analysis on primary tumor
332 samples with PROGENy[22,25], DoROthEA[24,25] and CytoSig[27], respectively. Scores were computed using
333 the *dorothea-py* and *progeny-py* packages. The top 1,000 target genes of the progeny model were used,

as recommended for single-cell data. For dorothea, only regulons of the highest confidence levels "A" and "B" were used. The cytosig signature matrix was obtained from the *data2intelligence/CytoSig* GitHub repository and used with the scoring function implemented in the *progeny-py* package. The methods were run with the options *num_perm=0, center=True, norm=True scale=True,* and *min_size=5.* No permutations were used, as we perform statistics in a separate step at the level of biological replicates. Pathway-, transcription factor-, and cytosig scores were then compared between *condition* and patient *group* (T *vs*. B *vs.* M *vs.* deserted) using an ordinary least-squares (OLS) linear model, as implemented in the statsmodels package [81]. Scores were aggregated into pseudobulk samples by computing the mean of each variable for each patient and cell-type. Samples consisting of less than 10 cells were discarded. For each variable, we fitted a model with the formulas ~ *condition* + *dataset* + *tumor_stage* or ~ *group* + *dataset* + *condition* + *tumor_stage*, respectively. Coefficients were obtained from the linear model and p-values calculated with the f-test. P-values were adjusted for multiple testing with the Benjamini-Hochberg procedure.

LIANA analysis

The ligand-receptor inference was computed using LIANA[29] framework to explore differences in cell-to-cell communication in primary tumor samples. The default *consensus* resource was used, which combines multiple expert-curated ligand-receptor databases, such as CellPhoneDB[82], CellChat[83], ICELLNET[84], connectomeDB2020, and CellTalkDB[85]. The *rank_aggregate* method was chosen combining ligand-receptor scores from multiple methods. Each method infers relevant interactions based on different assumptions, providing scores that reflect the interaction's strength and its specificity to particular cell identities. Our study focused on differences between conditions (tumor vs normal), treating patients as biological replicates. To identify differentially expressed in CCC events, we combined differential expression results for each gene into statistics of potentially downregulated ligand-receptor interactions. This analysis used adjusted p-values and the Wald test statistic. Interactions with low expression were filtered out when either a gene was expressed in less than 0.1 of cells, which is the default threshold. The fraction of cells expressing a signaling molecule was calculated as the mean of fractions per patient.

SCISSOR analysis

We used SCISSOR[86] to associate phenotypic data from bulk RNA-seq experiments with our single-cell data. TCGA mutation and gene expression data was obtained from the GDC portal, survival data from [86]. SCISSOR was run on primary tumor cells of each patient individually according to the SCISSOR tutorial using mutation data (logistic regression) and overall survival (cox-regression) as dependent variables. A grid search for the alpha-parameter was performed in $2^{-i/2}$ with $i \in [24, 23, \ldots, 2]$ and a cutoff parameter of 0.3. 62 of 674 samples with low overall cell count failed during SCISSOR's Seurat-preprocessing step and were excluded from the subsequent analysis. For each patient and cell-type, we

371 computed the fraction of *scissor+ cells* (i.e. positively associated with a mutation or worse survival),
372 *scissor- cells* (i.e. negatively associated), and *neutral cells* and added a pseudo-count of 0.01. A sample
373 was excluded from a cell-type if it contributed ≤10 cells. For each cell-type, we computed the log2-
374 ratio of scissor+ and scissor- cells as the mean fraction of scissor+ cells *vs.* the mean fraction of scissor-
375 cells. Significant differences were determined by comparing the fractions of scissor+ and scissor- cells
376 with a paired wilcoxon test with *zero_method="zsplit"* as implemented in the scipy package. P-values
377 were Benjamini-Hochberg-adjusted and considered significant at an FDR < 0.01.
378
379 ***Survival analysis***
380 Initially, gene lists were extracted from the patient groups according to their immune infiltration profiles
381 (B, desert, M and T). The gene lists consisted of differentially expressed genes, as returned from DESeq2
382 for the comparisons of each immune infiltration group with the rest, that were filtered using as thresholds
383 the adjusted p-values (< 0.05) and the logFold Change (greater than 0.5 or lower than -0.5).
384 Consequently, the gene lists were condensed into gene signatures expression values using the GSVA
385 (v1.50.5) R package. Once the gene signature expression values were calculated the TCGA COAD-
386 READ cohort (n = 629) was stratified into "high" and "low" expression, using as threshold the optimal
387 cutpoint as computed by the surv_cutpoint() function of the survminer (v0.4.9) R package. The survival
388 analysis was carried out using the survival (v3.7-0) R package and the Kaplan-Meier curves were plotted
389 using the ggsurvplot function of the survminer (v0.4.9) R package.
390
391 ***Quantification and statistical analysis***
392 Statistical analysis was performed using the statsmodels library in Python (scRNA-seq data) or
393 GraphPad Prism (flow cytometry and imaging data) using a linear model, t-test or wilcoxon test as
394 appropriate. Single cell-data were aggregated into pseudobulk samples by biological replicates.
395 Compositional analysis of cell-type fractions was performed using scCODA; survival analysis using the
396 Kaplan-Meier formula in R. P-values for untargeted analyses (DE genes, TFs, or pathways) were FDR-
397 adjusted. Significance levels and more details on the statistical tests are indicated in the figure captions.
398

399 **REFERENCES**

400

401 1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., and Jemal, A.
402 (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality
403 worldwide for 36 cancers in 185 countries. CA Cancer J Clin *74*, 229-263.
404 10.3322/caac.21834.

405 2. Akimoto, N., Ugai, T., Zhong, R., Hamada, T., Fujiyoshi, K., Giannakis, M., Wu, K., Cao, Y.,
406 Ng, K., and Ogino, S. (2021). Rising incidence of early-onset colorectal cancer - a call to
407 action. Nat Rev Clin Oncol *18*, 230-243. 10.1038/s41571-020-00445-1.

408 3. Cremolini, C., Loupakis, F., Antoniotti, C., Lupi, C., Sensi, E., Lonardi, S., Mezi, S.,
409 Tomasello, G., Ronzoni, M., Zaniboni, A., et al. (2015). FOLFOXIRI plus bevacizumab
410 versus FOLFIRI plus bevacizumab as first-line treatment of patients with metastatic colorectal
411 cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3
412 TRIBE study. Lancet Oncol *16*, 1306-1315. 10.1016/S1470-2045(15)00122-9.

413 4. Di Nicolantonio, F., Vitiello, P.P., Marsoni, S., Siena, S., Tabernero, J., Trusolino, L.,
414 Bernards, R., and Bardelli, A. (2021). Precision oncology in metastatic colorectal cancer -
415 from biology to medicine. Nat Rev Clin Oncol *18*, 506-525. 10.1038/s41571-021-00495-z.

416 5. Cancer Genome Atlas, N. (2012). Comprehensive molecular characterization of human colon
417 and rectal cancer. Nature *487*, 330-337. 10.1038/nature11252.

418 6. Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S.,
419 Kemberling, H., Wilt, C., Luber, B.S., et al. (2017). Mismatch-repair deficiency predicts
420 response of solid tumors to PD-1 blockade. Science. 10.1126/science.aan6733.

421 7. Galon, J., Costes, A., Sanchez-Cabo, F., Kirilovsky, A., Mlecnik, B., Lagorce-Pages, C.,
422 Tosolini, M., Camus, M., Berger, A., Wind, P., et al. (2006). Type, density, and location of
423 immune cells within human colorectal tumors predict clinical outcome. Science *313*, 1960-
424 1964. 10.1126/science.1129139.

425 8. Sahin, I.H., Ciombor, K.K., Diaz, L.A., Yu, J., and Kim, R. (2022). Immunotherapy for
426 Microsatellite Stable Colorectal Cancers: Challenges and Novel Therapeutic Avenues. Am
427 Soc Clin Oncol Educ Book *42*, 1-12. 10.1200/EDBK_349811.

428 9. Ros, J., Balconi, F., Baraibar, I., Saoudi Gonzalez, N., Salva, F., Tabernero, J., and Elez, E.
429 (2023). Advances in immune checkpoint inhibitor combination strategies for microsatellite
430 stable colorectal cancer. Front Oncol *13*, 1112276. 10.3389/fonc.2023.1112276.

431 10. Antoniotti, C., Rossini, D., Pietrantonio, F., Catteau, A., Salvatore, L., Lonardi, S., Boquet, I.,
432 Tamberi, S., Marmorino, F., Moretto, R., et al. (2022). Upfront FOLFOXIRI plus

433 bevacizumab with or without atezolizumab in the treatment of patients with metastatic
434 colorectal cancer (AtezoTRIBE): a multicentre, open-label, randomised, controlled, phase 2
435 trial. Lancet Oncol *23*, 876-887. 10.1016/S1470-2045(22)00274-1.

436 11. Hackl, H., Charoentong, P., Finotello, F., and Trajanoski, Z. (2016). Computational genomics
437 tools for dissecting tumour-immune cell interactions. Nat Rev Genet *17*, 441-458.
438 10.1038/nrg.2016.67.

439 12. Finotello, F., Rieder, D., Hackl, H., and Trajanoski, Z. (2019). Next-generation computational
440 tools for interrogating cancer immunity. Nat Rev Genet *20*, 724-746. 10.1038/s41576-019-
441 0166-7.

442 13. Angelova, M., Charoentong, P., Hackl, H., Fischer, M.L., Snajder, R., Krogsdam, A.M.,
443 Waldner, M.J., Bindea, G., Mlecnik, B., Galon, J., and Trajanoski, Z. (2015). Characterization
444 of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape
445 mechanisms and novel targets for immunotherapy. Genome biology *16*, 64. 10.1186/s13059-
446 015-0620-6.

447 14. Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H.,
448 and Trajanoski, Z. (2017). Pan-cancer immunogenomic analyses reveals genotype-
449 immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep
450 *18*, 248-262. 10.1101/056101.

451 15. Salcher, S., Sturm, G., Horvath, L., Untergasser, G., Kuempers, C., Fotakis, G., Panizzolo, E.,
452 Martowicz, A., Trebo, M., Pall, G., et al. (2022). High-resolution single-cell atlas reveals
453 diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. Cancer cell
454 *40*, 1503-1520 e1508. 10.1016/j.ccell.2022.10.008.

455 16. Oliveira, G., Stromhaug, K., Klaeger, S., Kula, T., Frederick, D.T., Le, P.M., Forman, J.,
456 Huang, T., Li, S., Zhang, W., et al. (2021). Phenotype, specificity and avidity of antitumour
457 CD8(+) T cells in melanoma. Nature *596*, 119-125. 10.1038/s41586-021-03704-y.

458 17. He, X., Wang, L., Riedel, H., Wang, K., Yang, Y., Dinu, C.Z., and Rojanasakul, Y. (2017).
459 Mesothelin promotes epithelial-to-mesenchymal transition and tumorigenicity of human lung
460 cancer and mesothelioma cells. Molecular cancer *16*, 63. 10.1186/s12943-017-0633-8.

461 18. Francart, M.E., Vanwynsberghe, A.M., Lambert, J., Bourcy, M., Genna, A., Ancel, J., Perez-
462 Boza, J., Noel, A., Birembaut, P., Struman, I., et al. (2020). Vimentin prevents a miR-
463 dependent negative regulation of tissue factor mRNA during epithelial-mesenchymal
464 transitions and facilitates early metastasis. Oncogene *39*, 3680-3692. 10.1038/s41388-020-
465 1244-1.

466 19. Kubala, M.H., Punj, V., Placencio-Hickok, V.R., Fang, H., Fernandez, G.E., Sposto, R., and
467 DeClerck, Y.A. (2018). Plasminogen Activator Inhibitor-1 Promotes the Recruitment and
468 Polarization of Macrophages in Cancer. Cell Rep *25*, 2177-2191 e2177.
469 10.1016/j.celrep.2018.10.082.

470 20. Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt,
471 B., Buerger, H., Bulk, E., et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin
472 beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*,
473 8031-8041. 10.1038/sj.onc.1206928.

474 21. Qi, L., Liu, F., Zhang, F., Zhang, S., Lv, L., Bi, Y., and Yu, Y. (2018). lncRNA NEAT1
475 competes against let-7a to contribute to non-small cell lung cancer proliferation and
476 metastasis. Biomed Pharmacother *103*, 1507-1515. 10.1016/j.biopha.2018.04.053.

477 22. Schubert, M., Klinger, B., Klunemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J.,
478 Bluthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling
479 footprints in cancer gene expression. Nature communications *9*, 20. 10.1038/s41467-017-
480 02391-6.

481 23. Jiang, G., Shi, L., Zheng, X., Zhang, X., Wu, K., Liu, B., Yan, P., Liang, X., Yu, T., Wang,
482 Y., and Cai, X. (2020). Androgen receptor affects the response to immune checkpoint therapy
483 by suppressing PD-L1 in hepatocellular carcinoma. Aging *12*, 11466-11484.
484 10.18632/aging.103231.

485 24. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019).
486 Benchmark and integration of resources for the estimation of human transcription factor
487 activities. Genome research *29*, 1363-1375. 10.1101/gr.240663.118.

488 25. Holland, C.H., Tanevski, J., Perales-Paton, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin,
489 B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020). Robustness and applicability of
490 transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol *21*,
491 36. 10.1186/s13059-020-1949-z.

492 26. Peng, S.L. (2008). Foxo in the immune system. Oncogene *27*, 2337-2344.
493 10.1038/onc.2008.26.

494 27. Jiang, P., Zhang, Y., Ru, B., Yang, Y., Vu, T., Paul, R., Mirza, A., Altan-Bonnet, G., Liu, L.,
495 Ruppin, E., et al. (2021). Systematic investigation of cytokine signaling activity at the tissue
496 and single-cell levels. Nat Methods *18*, 1181-1191. 10.1038/s41592-021-01274-5.

497 28. Yang, W.C., Hwang, Y.S., Chen, Y.Y., Liu, C.L., Shen, C.N., Hong, W.H., Lo, S.M., and
498 Shen, C.R. (2017). Interleukin-4 Supports the Suppressive Immune Responses Elicited by
499 Regulatory T Cells. Front Immunol *8*, 1508. 10.3389/fimmu.2017.01508.

500 29. Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P.L., Nagai, J.S., Boys, C.,
501 Ramirez Flores, R.O., Kim, H., Szalai, B., Costa, I.G., et al. (2022). Comparison of methods
502 and resources for cell-cell communication inference from single-cell RNA-seq data. Nat.
503 Commun. 13, 3224. 10.1038/s41467-022-30755-0.

504 30. Pylaeva, E., and Jablonska, J. (2023). Sellhi neutrophils: a key to triumph in cancer
505 therapy. *Signal Transduction and Targeted Therapy* **8**, 421 10.1038/s41392-023-01671-6.

506 31. Marzio, R., Mauël, J., and Betz-Corradin, S. (1999). CD69 and regulation of the
507 immune function. Immunopharmacology and Immunotoxicology 21, 565–582.
508 10.3109/08923979909007126.

509 32. Del Prete, A., Martínez-Muñoz, L., Mazzon, C., Toffali, L., Sozio, F., Za, L., Bosisio,
510 D., Gazzurelli, L., Salvi, V., Tiberio, L., et al. (2017). The atypical receptor CCRL2 is
511 required for CXCR2-dependent neutrophil recruitment and tissue damage. Blood 130, 1223–
512 1234. 10.1182/blood-2017-04-777680.

513 29. Motz, G.T., and Coukos, G. (2011). The parallel lives of angiogenesis and
514 immunosuppression: cancer and other tales. Nat Rev Immunol *11*, 702-711. 10.1038/nri3064.

515 30. La Fleur, L., Botling, J., He, F., Pelicano, C., Zhou, C., He, C., Palano, G., Mezheyeuski, A.,
516 Micke, P., Ravetch, J.V., et al. (2021). Targeting MARCO and IL37R on Immunosuppressive
517 Macrophages in Lung Cancer Blocks Regulatory T Cells and Supports Cytotoxic Lymphocyte
518 Function. Cancer research *81*, 956-967. 10.1158/0008-5472.CAN-20-1885.

519 31. Yin, Y., Xu, X., Tang, J., Zhang, W., Zhangyuan, G., Ji, J., Deng, L., Lu, S., Zhuo, H., and
520 Sun, B. (2018). CD97 Promotes Tumor Aggressiveness Through the Traditional G Protein-
521 Coupled Receptor-Mediated Signaling in Hepatocellular Carcinoma. Hepatology *68*, 1865-
522 1878. 10.1002/hep.30068.

523 32. Shojaei, F., Scott, N., Kang, X., Lappin, P.B., Fitzgerald, A.A., Karlicek, S., Simmons, B.H.,
524 Wu, A., Lee, J.H., Bergqvist, S., and Kraynov, E. (2012). Osteopontin induces growth of
525 metastatic tumors in a preclinical model of non-small lung cancer. J Exp Clin Cancer Res *31*,
526 26. 10.1186/1756-9966-31-26.

527 33. Zhang, J., Takahashi, K., Takahashi, F., Shimizu, K., Ohshita, F., Kameda, Y., Maeda, K.,
528 Nishio, K., and Fukuchi, Y. (2001). Differential osteopontin expression in lung cancer. Cancer
529 Lett *171*, 215-222. 10.1016/s0304-3835(01)00607-3.

530 34. Meng, J., Jiang, Y.Z., Zhao, S., Tao, Y., Zhang, T., Wang, X., Zhang, Y., Sun, K., Yuan, M.,
531 Chen, J., et al. (2022). Tumor-derived Jagged1 promotes cancer progression through immune
532 evasion. Cell Rep *38*, 110492. 10.1016/j.celrep.2022.110492.

533 35. Mikucki, M.E., Fisher, D.T., Matsuzaki, J., Skitzki, J.J., Gaulin, N.B., Muhitch, J.B., Ku,
534 A.W., Frelinger, J.G., Odunsi, K., Gajewski, T.F., et al. (2015). Non-redundant requirement
535 for CXCR3 signalling during tumoricidal T-cell trafficking across tumour vascular
536 checkpoints. Nature communications *6*, 7458. 10.1038/ncomms8458.

537 36. Choi, H.Y., Jo, H., Zhao, X., Hoadley, K.A., Newman, S., Holt, J., Hayward, M.C., Love,
538 M.I., Marron, J.S., and Hayes, D.N. (2021). SCISSOR: a framework for identifying structural
539 changes in RNA transcripts. Nature communications *12*, 286. 10.1038/s41467-020-20593-3.

540 37. Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H.,
541 and Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-
542 Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell
543 Rep *18*, 248-262. 10.1016/j.celrep.2016.12.019.

544 38. Chen S, Chen M. Schwann cells promote the migration and invasion of colorectal cancer cells
545 via the activated NF-κB/IL-8 axis in the tumor microenvironment. Front Oncol. 2022 Nov
546 17;12:1026670. doi: 10.3389/fonc.2022.1026670.

547 39. Wellenstein, M.D., and de Visser, K.E. (2018). Cancer-Cell-Intrinsic Mechanisms Shaping the
548 Tumor Immune Landscape. Immunity *48*, 399-416. 10.1016/j.immuni.2018.03.004.

549 40. Patel, A.J., Richter, A., Drayson, M.T., and Middleton, G.W. (2020). The role of B
550 lymphocytes in the immuno-biology of non-small-cell lung cancer. Cancer immunology,
551 immunotherapy : CII *69*, 325-342. 10.1007/s00262-019-02461-2.

552 41. Li, T., Li, X., Zamani, A., Wang, W., Lee, C.N., Li, M., Luo, G., Eiler, E., Sun, H., Ghosh, S.,
553 et al. (2020). c-Rel Is a Myeloid Checkpoint for Cancer Immunotherapy. Nat Cancer *1*, 507-
554 517. 10.1038/s43018-020-0061-3.

555 42. Fridman, W.H., Meylan, M., Petitprez, F., Sun, C.M., Italiano, A., and Sautès-Fridman, C.
556 (2022). B cells and tertiary lymphoid structures as determinants of tumour immune contexture
557 and clinical outcome. Nat Rev Clin Oncol *19*, 441-457. 10.1038/s41571-022-00619-z.

558 43. Dong, Z.Y., Zhang, J.T., Liu, S.Y., Su, J., Zhang, C., Xie, Z., Zhou, Q., Tu, H.Y., Xu, C.R.,
559 Yan, L.X., et al. (2017). EGFR mutation correlates with uninflamed phenotype and weak
560 immunogenicity, causing impaired response to PD-1 blockade in non-small cell lung cancer.
561 Oncoimmunology *6*, e1356145. 10.1080/2162402x.2017.1356145.

562 44. Pinto, R., Petriella, D., Lacalamita, R., Montrone, M., Catino, A., Pizzutilo, P., Botticella,
563 M.A., Zito, F.A., Del Bene, G., Zonno, A., et al. (2019). KRAS-Driven Lung Adenocarcinoma
564 and B Cell Infiltration: Novel Insights for Immunotherapy. Cancers (Basel) *11*.
565 10.3390/cancers11081145.

566 45. Leader, A.M., Grout, J.A., Maier, B.B., Nabet, B.Y., Park, M.D., Tabachnikova, A., Chang,
567 C., Walker, L., Lansky, A., Le Berichel, J., et al. (2021). Single-cell analysis of human non-
568 small cell lung cancer lesions refines tumor classification and patient stratification. Cancer
569 Cell *39*, 1594-1609 e1512. 10.1016/j.ccell.2021.10.009.

570 57. Madissoon, E., Wilbrey-Clark, A., Miragaia, R.J., Saeb-Parsy, K., Mahbubani, K.T.,
571 Georgakopoulos, N., Harding, P., Polanski, K., Huang, N., Nowicki-Osuch, K., et al. (2019).
572 scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold
573 preservation. Genome Biol *21*, 1. 10.1186/s13059-019-1906-x.

574 58. Reyfman, P.A., Walter, J.M., Joshi, N., Anekalla, K.R., McQuattie-Pimentel, A.C., Chiu, S.,
575 Fernandez, R., Akbarpour, M., Chen, C.I., Ren, Z., et al. (2019). Single-Cell Transcriptomic
576 Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. Am
577 J Respir Crit Care Med *199*, 1517-1536. 10.1164/rccm.201712-2410OC.

578 59. Vieira Braga, F.A., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M., Brouwer, S.,
579 Gomes, T., Hesse, L., Jiang, J., et al. (2019). A cellular census of human lungs identifies novel
580 cell states in health and in asthma. Nature medicine *25*, 1153-1163. 10.1038/s41591-019-
581 0468-5.

582 60. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G.,
583 Raby, B.A., DeIuliis, G., Januszyk, M., et al. (2020). Single-cell RNA-seq reveals ectopic and
584 aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv *6*, eaba1983.
585 10.1126/sciadv.aba1983.

586 61. Habermann, A.C., Gutierrez, A.J., Bui, L.T., Yahn, S.L., Winters, N.I., Calvi, C.L., Peter, L.,
587 Chung, M.I., Taylor, C.J., Jetter, C., et al. (2020). Single-cell RNA sequencing reveals
588 profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci
589 Adv *6*, eaba1972. 10.1126/sciadv.aba1972.

590 62. Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S.,
591 Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from
592 single-cell RNA sequencing. Nature *587*, 619-625. 10.1038/s41586-020-2922-4.

593 63. Picelli, S., Björklund, Å., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013).
594 Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods *10*,
595 1096-1098. 10.1038/nmeth.2639.

596 64. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz,
597 D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied
598 to embryonic stem cells. Cell *161*, 1187-1201. 10.1016/j.cell.2015.04.044.

599 65. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas,
600 A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression
601 Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.
602 10.1016/j.cell.2015.05.002.

603 66. Maier, B., Leader, A.M., Chen, S.T., Tung, N., Chang, C., LeBerichel, J., Chudnovskiy, A.,
604 Maskey, S., Walker, L., Finnigan, J.P., et al. (2020). A conserved dendritic-cell regulatory
605 program limits antitumour immunity. Nature *580*, 257-262. 10.1038/s41586-020-2134-y.

606 67. Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G.S., Putman, T.E.,
607 Ainscough, B.J., Griffith, O.L., et al. (2016). High-performance web services for querying
608 gene and variant annotation. Genome Biol *17*, 91. 10.1186/s13059-016-0953-9.

609 68. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic
610 harmonization and annotation of single-cell transcriptomics data with deep generative models.
611 Mol Syst Biol *17*, e9620. 10.15252/msb.20209620.

612 69. Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K.,
613 Jayasuriya, M., Mehlman, E., Langevin, M., et al. (2022). A Python library for probabilistic
614 analysis of single-cell omics data. Nat Biotechnol *40*, 163-166. 10.1038/s41587-021-01206-w.

615 70. Luecken, M.D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F.,
616 Strobl, D.C., Zappia, L., Dugas, M., Colome-Tatche, M., and Theis, F.J. (2022).
617 Benchmarking atlas-level data integration in single-cell genomics. Nat Methods *19*, 41-50.
618 10.1038/s41592-021-01336-8.

619 71. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative
620 modeling for single-cell transcriptomics. Nat Methods *15*, 1053-1058. 10.1038/s41592-018-
621 0229-2.

622 72. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene
623 expression data analysis. Genome Biol *19*, 15. 10.1186/s13059-017-1382-0.

624 73. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and
625 Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP.
626 Nat Biotechnol. 10.1038/nbt.4314.

627 74. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing
628 well-connected communities. Sci Rep *9*, 5233. 10.1038/s41598-019-41695-z.

629 75. Schupp, J.C., Adams, T.S., Cosme, C., Jr., Raredon, M.S.B., Yuan, Y., Omote, N., Poli, S.,
630 Chioccioli, M., Rose, K.A., Manning, E.P., et al. (2021). Integrated Single-Cell Atlas of
631 Endothelial Cells of the Human Lung. Circulation *144*, 286-302.
632 10.1161/CIRCULATIONAHA.120.052318.

633 76. Sikkema, L., Strobl, D., Zappia, L., Madissoon, E., Markov, N.S., Zaragosi, L., Ansari, M.,
634 Arguel, M., Apperloo, L., Bécavin, C., et al. (2022). An integrated cell atlas of the human lung
635 in health and disease. bioRxiv. 10.1101/2022.03.10.483747.

636 77. Buttner, M., Ostner, J., Muller, C.L., Theis, F.J., and Schubert, B. (2021). scCODA is a
637 Bayesian model for compositional single-cell data analysis. Nat Commun *12*, 6876.
638 10.1038/s41467-021-27150-6.

639 78. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
640 dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-
641 8.

642 79. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R.,
643 Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-
644 cell differential expression. Nature communications *12*, 5692. 10.1038/s41467-021-25960-2.

645 80. Ignatiadis, N., Klaus, B., Zaugg, J.B., and Huber, W. (2016). Data-driven hypothesis
646 weighting increases detection power in genome-scale multiple testing. Nat Methods *13*, 577-
647 580. 10.1038/nmeth.3885.

648 81. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with
649 Python. Proceedings of the Python in Science Conference. 10.25080/majora-92bf1922-011.

650 82. Efremova, M., Vento-Tormo, M., Teichmann, S. A., & others. (2020). CellPhoneDB:
651 Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor
652 complexes. Nat. Protoc. 15, 1484–1506. 10.1038/s41596-020-0292-x.

653 83. Jin, S., Guerrero-Juarez, C. F., Zhang, L., & others. (2021). Inference and analysis of cell-cell
654 communication using CellChat. Nat. Commun. 12, 1088. 10.1038/s41467-021-21246-9.

655 84. Noël, F., Massenet-Regad, L., Carmi-Levy, I., & others. (2021). Dissection of intercellular
656 communication using the transcriptome-based framework ICELLNET. Nat. Commun. 12,
657 1089.10.1038/s41467-021-21244-x.

658 85. Shao, X., Liao, J., Li, C., Lu, X., Cheng, J., & Fan, X. (2021). CellTalkDB: A manually
659 curated database of ligand-receptor interactions in humans and mice. Brief. Bioinform. 22,
660 bbaa269.10.1093/bib/bbaa269.

661 86. Sun, D., Guan, X., Moran, A.E., Wu, L.Y., Qian, D.Z., Schedin, P., Dai, M.S., Danilov, A.V.,
662 Alumkal, J.J., Adey, A.C., et al. (2022). Identifying phenotype-associated subpopulations by
663 integrating bulk and single-cell sequencing data. Nat Biotechnol *40*, 527-538.
664 10.1038/s41587-021-01091-3.
665

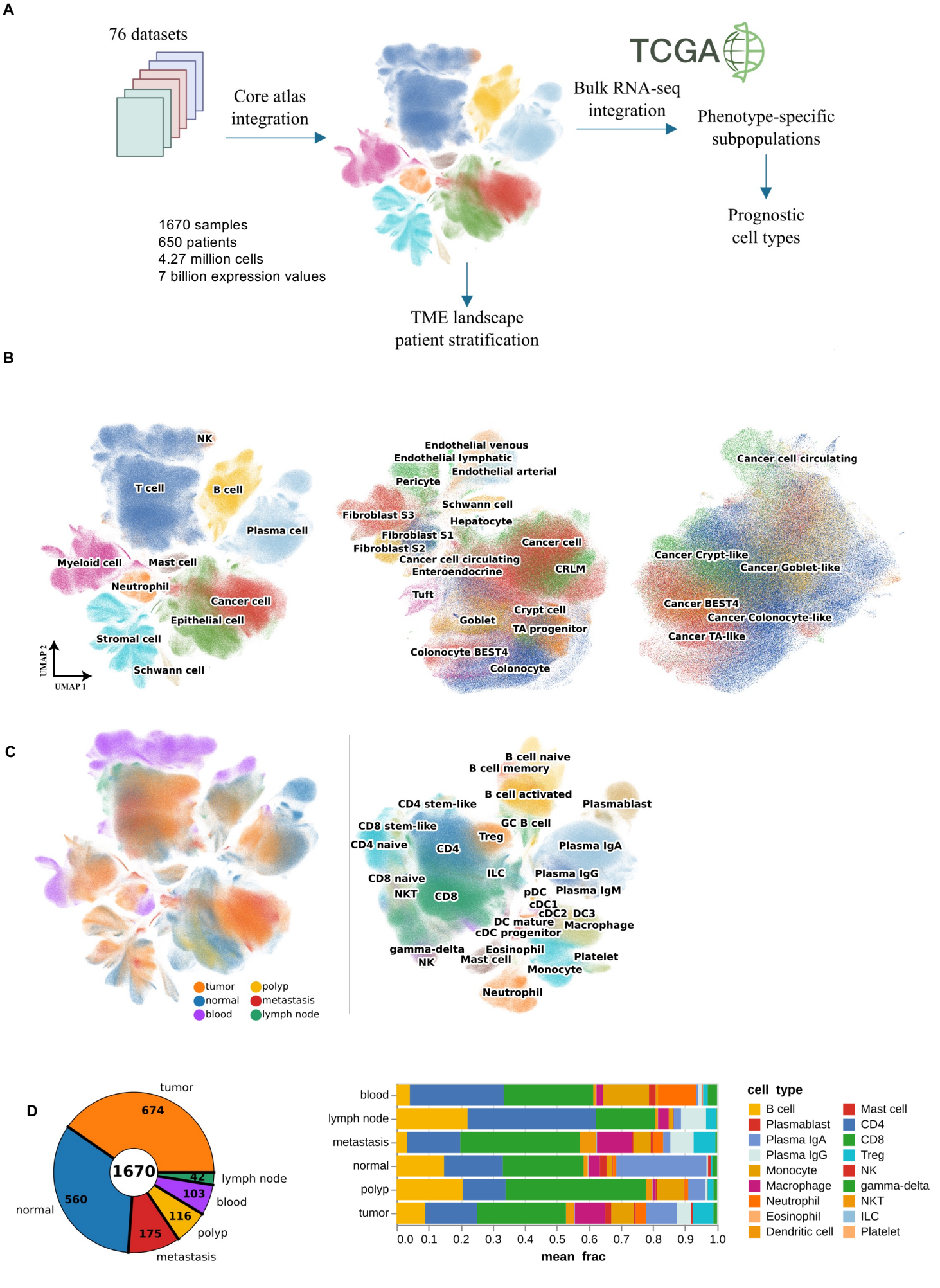
A
76 datasets
Core atlas integration
1670 samples
650 patients
4.27 million cells
7 billion expression values
TME landscape patient stratification
TCGA
Bulk RNA-seq integration
Phenotype-specific subpopulations
Prognostic cell types
B
NK
T cell
B cell
Plasma cell
Myeloid cell
Mast cell
Neutrophil
Cancer cell
Epithelial cell
Stromal cell
Schwann cell
UMAP 2
UMAP 1
Endothelial venous
Endothelial lymphatic
Endothelial arterial
Pericyte
Schwann cell
Fibroblast S3
Hepatocyte
Fibroblast S1
Fibroblast S2
Cancer cell
Cancer cell circulating
Enteroendocrine
CRLM
Tuft
Crypt cell
Goblet
TA progenitor
Colonocyte BEST4
Colonocyte
Cancer cell circulating
Cancer Crypt-like
Cancer Goblet-like
Cancer BEST4
Cancer Colonocyte-like
Cancer TA-like
C
tumor
polyp
normal
metastasis
blood
lymph node
B cell naive
B cell memory
B cell activated
Plasmablast
CD4 stem-like
CD8 stem-like
GC B cell
Treg
CD4 naive
CD4
Plasma IgA
ILC
Plasma IgG
CD8 naive
NKT
CD8
pDC
Plasma IgM
cDC1
cDC2
DC3
DC mature
Macrophage
cDC progenitor
gamma-delta
Eosinophil
NK
Mast cell
Platelet
Monocyte
Neutrophil
D
tumor
674
1670
42
lymph node
103
blood
560
normal
116
175
polyp
metastasis
blood
lymph node
metastasis
normal
polyp
tumor
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
mean frac
cell type
B cell
Plasmablast
Plasma IgA
Plasma IgG
Monocyte
Macrophage
Neutrophil
Eosinophil
Dendritic cell
Mast cell
CD4
CD8
Treg
NK
gamma-delta
NKT
ILC
Platelet

Figure 1:

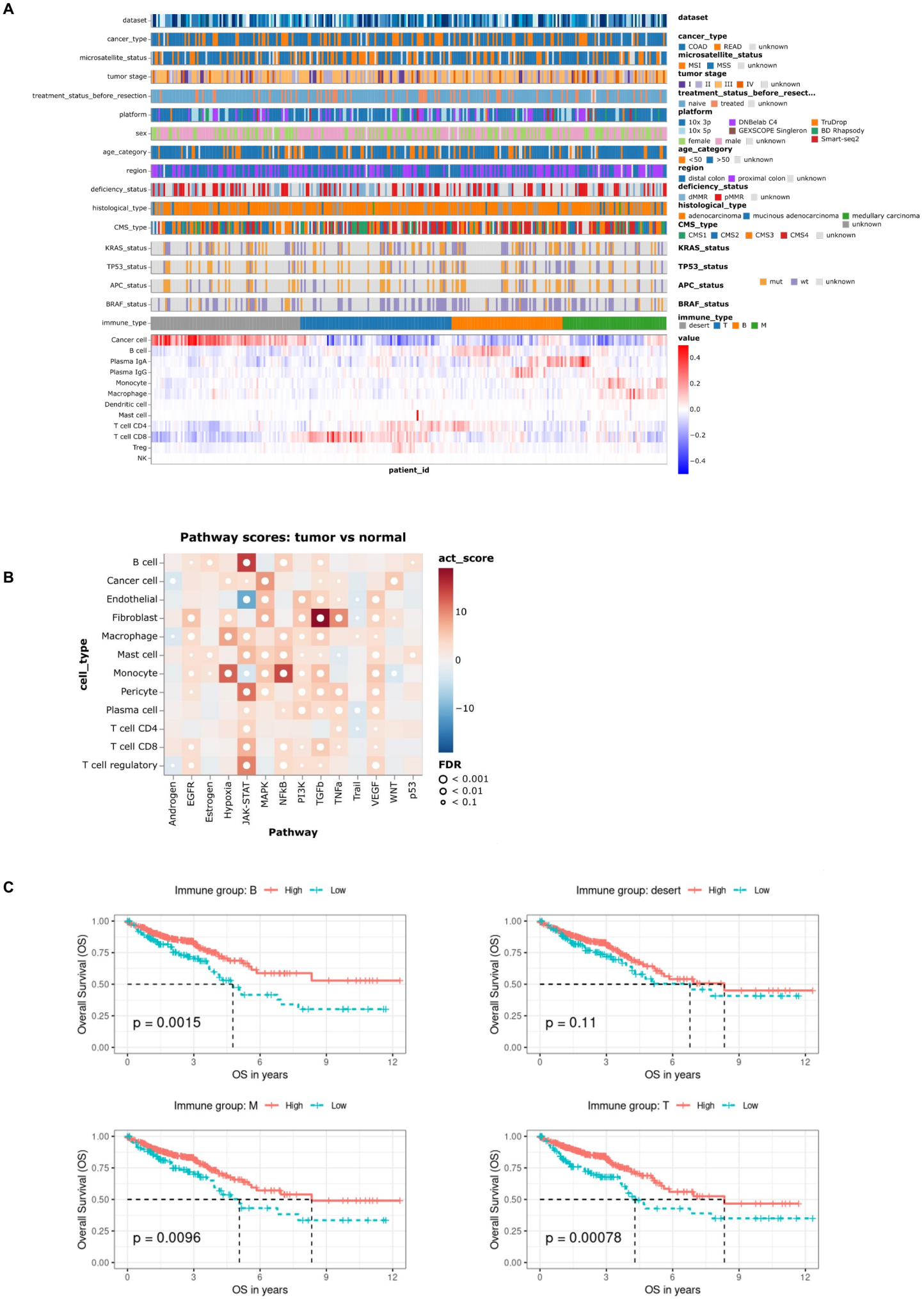
A
dataset
cancer_type
microsatellite_status
tumor stage
treatment_status_before_resection
platform
sex
age_category
region
deficiency_status
histological_type
CMS_type
KRAS_status
TP53_status
APC_status
BRAF_status
immune_type
Cancer cell
B cell
Plasma IgA
Plasma IgG
Monocyte
Macrophage
Dendritic cell
Mast cell
T cell CD4
T cell CD8
Treg
NK
patient_id
COAD READ unknown
MSI MSS unknown
I II III IV unknown
naive treated unknown
10x 3p DNBelab C4 TruDrop
10x 5p GEXSCOPE Singleron BD Rhapsody
female male unknown Smart-seq2
<50 >50 unknown
distal colon proximal colon unknown
dMMR pMMR unknown
adenocarcinoma mucinous adenocarcinoma medullary carcinoma
unknown
CMS1 CMS2 CMS3 CMS4 unknown
mut wt unknown
desert T B M
value
0.4
0.2
0.0
−0.2
−0.4
B
Pathway scores: tumor vs normal
cell_type
B cell
Cancer cell
Endothelial
Fibroblast
Macrophage
Mast cell
Monocyte
Pericyte
Plasma cell
T cell CD4
T cell CD8
T cell regulatory
Androgen
EGFR
Estrogen
Hypoxia
JAK-STAT
MAPK
NFkB
PI3K
TGFb
TNFa
Trail
VEGF
WNT
p53
Pathway
act_score
10
0
−10
FDR
< 0.001
< 0.01
< 0.1
C
Immune group: B High Low
Immune group: desert High Low
Immune group: M High Low
Immune group: T High Low
Overall Survival (OS)
OS in years
p = 0.0015
p = 0.11
p = 0.0096
p = 0.00078

Figure 2

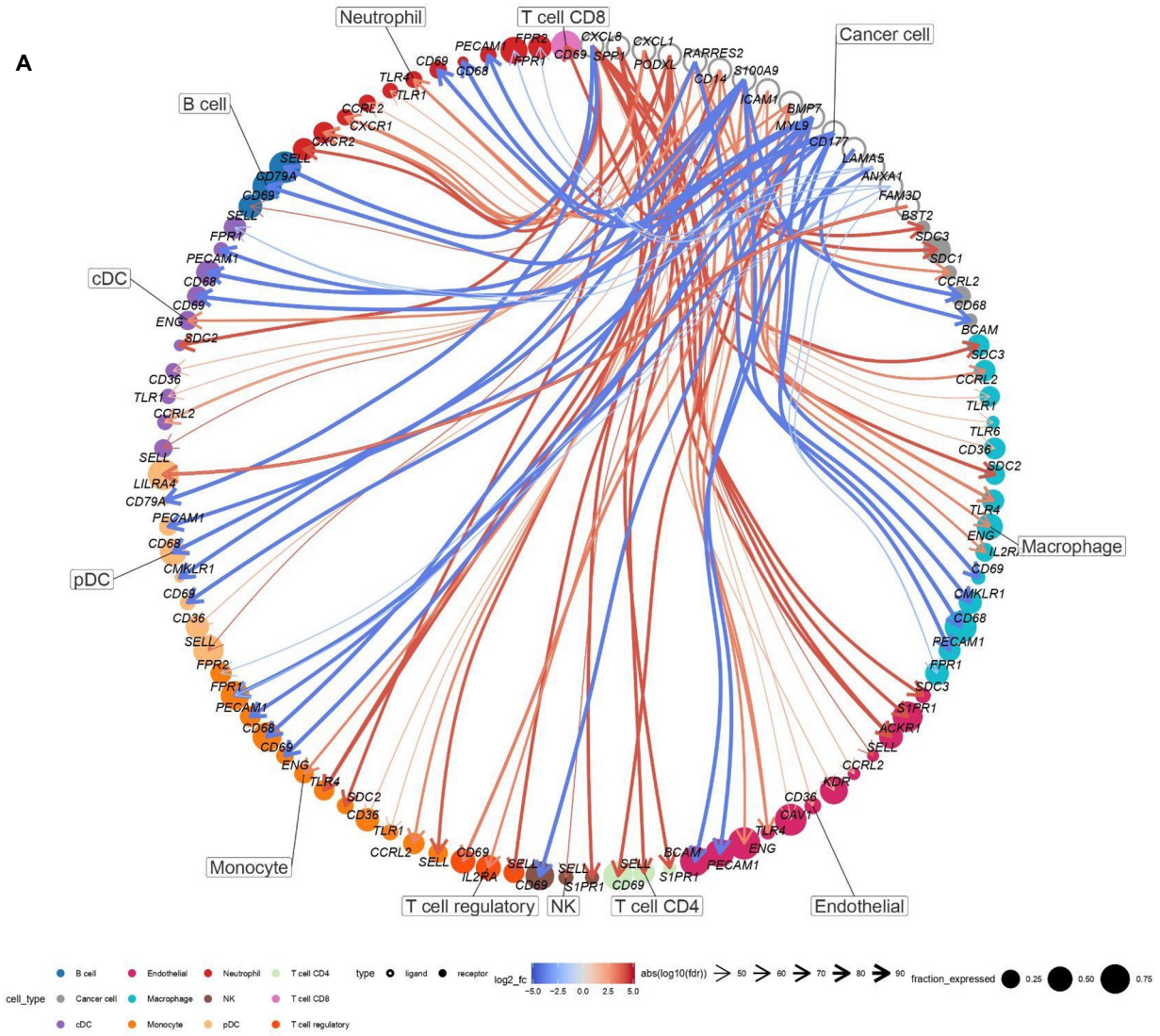
A
Neutrophil
T cell CD8
Cancer cell
B cell
cDC
pDC
Macrophage
Monocyte
T cell regulatory
NK
T cell CD4
Endothelial
cell_type
B cell
Cancer cell
cDC
Endothelial
Macrophage
Monocyte
Neutrophil
NK
pDC
T cell CD4
T cell CD8
T cell regulatory
type
ligand
receptor
log2_fc
-5.0 -2.5 0.0 2.5 5.0
abs(log10(fdr))
50 60 70 80 90
fraction_expressed
0.25 0.50 0.75

Figure 3

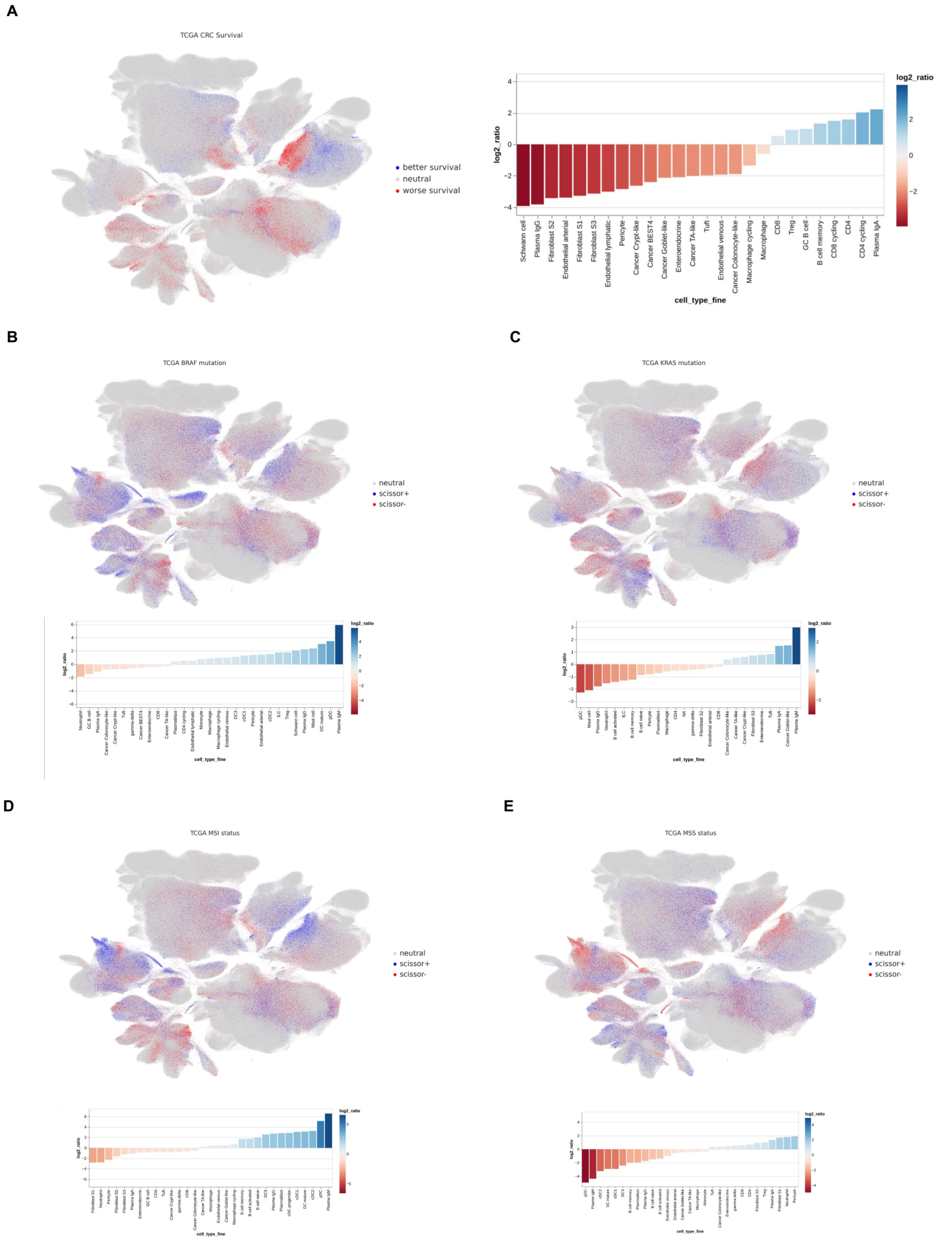
A
TCGA CRC Survival
better survival
neutral
worse survival
log2_ratio
cell_type_fine
Schwann cell
Plasma IgG
Fibroblast S2
Endothelial arterial
Fibroblast S1
Fibroblast S3
Endothelial lymphatic
Pericyte
Cancer Crypt-like
Cancer BEST4
Cancer Goblet-like
Enteroendocrine
Cancer TA-like
Tuft
Endothelial venous
Cancer Colonocyte-like
Macrophage cycling
Macrophage
CD8
Treg
GC B cell
B cell memory
CD8 cycling
CD4
CD4 cycling
Plasma IgA
B
TCGA BRAF mutation
neutral
scissor+
scissor-
C
TCGA KRAS mutation
neutral
scissor+
scissor-
D
TCGA MSI status
neutral
scissor+
scissor-
E
TCGA MSS status
neutral
scissor+
scissor-

Figure 4