Mapping Cell-to-cell Interactions from Spatially Resolved Transcriptomics Data

James Zhu[1,*], Yunguan Wang[1,2,3*], Woo Yong Chang[1,*], Alicia Malewska[4], Fabiana Napolitano[5], Jeffrey C. Gahan[4], Nisha Unni[6], Min Zhao[7], Fangjiang Wu[1], Lauren Yue[1], Lei Guo[1], Zhuo Zhao[8], Danny Z. Chen[8], Raquibul Hannan[9], Siyuan Zhang[7], Guanghua Xiao[1,10], Ping Mu[11,12], Ariella B. Hanker[5], Douglas Strand[4], Carlos L. Arteaga[5], Neil Desai[9], Xinlei Wang[13,14,+], Yang Xie[1,10,+], Tao Wang[1,+]

1 Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
2 Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
3 Department of Pediatrics, University of Cincinnati, OH, 45221, USA.
4 Department of Urology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
5 Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
6 Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
7 Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
8 Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA.
9 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
10 Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
11 Department of Molecular Biology, UT Southwestern Medical Center, Dallas, TX, 75390, USA.
12 Hamon Center for Regenerative Science and Medicine, UT Southwestern Medical Center, Dallas, TX, 75390, USA.
13 Department of Mathematics, University of Texas at Arlington, Arlington, TX, 76019, USA.
14 Center for Data Science Research and Education, College of Science, University of Texas at Arlington, Arlington, TX, 76019, USA.
* Co-first authors
+ Co-corresponding authors: Xinlei Wang, xinlei.wang@uta.edu; Yang Xie, yang.xie@utsouthwestern.edu; Tao Wang, tao.wang@utsouthwestern.edu

Short title: Inferring cell-to-cell interaction from spatial transcriptomics data

Keywords: spatially resolved transcriptomics, cell-to-cell interaction, spacia, MERSCOPE, GeoMX

ABSTRACT

Cell-cell communication (CCC) is essential to how life forms, develops and functions. However, accurate, high-throughput mapping of how expression of all genes in one cell affects expression of all genes in another cell has been bottlenecked by under-developed experimental techniques and inadequate analytical designs. Here, we introduce a Bayesian multi-instance learning framework, spacia, to detect CCC from emerging spatially resolved transcriptomics (SRT) data by uniquely exploiting their spatial modality. We highlight spacia's power to overcome fundamental limitations of popular single-cell RNA sequencing-based tools for inference of CCC, which lose single-cell resolution of CCCs and suffer from high false positive rates. Spacia unveiled how various types of cells in the tumor microenvironment differentially contribute to Epithelial-Mesenchymal Transition and lineage plasticity in tumor cells in a prostate cancer MERSCOPE dataset. We deployed spacia in a set of pan-cancer MERSCOPE datasets and derived a signature for measuring the impact of *PDL1* on receiving cells from PDL1-positive sending cells. We demonstrated that this signature is associated with patient survival and response to immune checkpoint inhibitor treatments in 3,354 patients. Overall, spacia represents a notable step in advancing quantitative theories of cellular communications.

INTRODUCTION

Various types of cells form complex structures and communication networks in the tissue micro-environment, and the signaling between these cells is central to normal organ development and diseased physiological processes. Elucidating cell-cell communication (CCC) in the tissue microenvironment in different biological systems is of vital importance. Many experimental and informatics approaches have attempted to address this question (*1–4*), and one major school of informatics approaches infer CCCs based on single-cell RNA-sequencing (scRNA-seq) data, such as CellChat (*5*), NicheNet (*6*), CellphoneDB (*7*), NATMI (*8*), SingleCellSignalR (*9*), *etc*. Despite their popularity, these methods suffer from a number of significant caveats, due to the limited information provided by scRNA-seq and improper designs of the underpinning models.

To begin with, most of these tools only infer interactions between cell types rather than interactions at the single-cell level, thus losing single cell resolution. Secondly, CCC is usually context-specific, and the common approach of mapping the data to pre-defined interaction pathway databases, regardless of the cellular context, inevitably results in low resolution and low sensitivity in the detection of true CCCs in the specific tissue sample of interest. Lastly, most of these tools rely on the co-expression of ligand-receptor gene pairs in signal-sending and receiving cells to claim detection of CCC. However, the expression of the receptor gene itself is not necessarily impacted by the expression of the ligand, which calls into question the fundamental rationale of such approaches. Rather, SRTs can only capture the alterations of downstream target genes' RNA expression in the receiving cells that are influenced by ligand signaling from the sending cells.

We hypothesize that effectively addressing CCCs requires examining the interacting cells in both their expression space and their physical space. Fortunately, emerging spatially resolved transcriptomics (SRT) technologies provide the necessary data to explore this possibility. Such SRT technologies include 10X Visium, Slide-Seq (*10*, *11*), Seq-scope (*12*), XYZeq (*13*), CODEX (Akoya Biosciences), CosMx (Nanostring), MERSCOPE (Vizgen), *etc*. First, these SRTs, especially those that achieve single cell spatial resolution, allow researchers to pinpoint each single cell with its spatial location and identify the other cells within its neighborhood. It is thus feasible to detect interactions between each pair of single cells without having to aggregate to cell types. In addition, since SRTs enable examining the expression profiles of all possible pairs of single cells, yielding very rich data that reveal ongoing CCCs, we can apply a much more data-driven approach to infer CCCs, overcoming the constraint of relying on previously known pathways. Furthermore, the high-dimensional and multi-modal information contained in these cell pair data also enables the modeling of downstream target genes in the receiving cells in detail, such as how their expression varies as a function of signals from sending cells. Therefore, SRTs overcome all major caveats encountered in mapping of CCCs from scRNA-seq data.

Although SRTs in principle provide the data necessary for accurate mapping of CCCs, these methods generally capture thousands of cells at a minimum, with information on each cell's location and expression of at least a few hundred genes. Such highly complex data present a grand challenge for analytics. To address this daunting task, we present spacia, a Bayesian multi-instance learning framework, to detect the interaction between signal-sending and receiving cells by reconciling the spatial and transcriptomics modalities. Importantly, cell-cell communications happen in a variety of manners. As reviewed by Armingol *et al (2)*, there are four major types of cell-cell communications: autocrine, paracrine, juxtacrine, and endocrine. The first three types of communication naturally require cells to be in close proximity. In contrast, endocrine interactions occur over long distances through systemic circulation, and it is not feasible to track such CCCs by SRT. Therefore, spacia considers the types of cell-cell communications that require the interacting cells to be closely localized, by leveraging the spatial modality of SRTs.

With spacia, we showed that data from modern SRTs already allow sophisticated analyses, more than merely tracking different types of cells in their spatial context. Spacia incorporates the unique spatial location information of SRT and enables precise and single-cell resolution detection of CCC, a capability not provided by scRNA-seq-based counterparts. When applied to a prostate MERSCOPE dataset, spacia revealed how stromal/immune cells contribute to EMT and lineage plasticity in prostate cancer cells. When applied to a pan-cancer panel of MERSCOPE datasets, spacia revealed a signature for measuring the impact of *PDL1* signaling on various types of immune cells, including $CD8^+$ T cells. We demonstrated that this spacia-derived PDL1-CD8 signature is predictive of patient survival and treatment response to immune checkpoint inhibitor treatments in 3,354 patients.

RESULTS

SRTs Reveal CCCs in the Tissue Microenvironment

While SRTs have prospered and some are even commercialized, there are still many technical challenges associated with modern SRT technologies such as shallow coverage for each gene. Cell typing within the spatial context by SRT is mostly feasible, but it remains unclear for the whole field whether current SRT technologies allow enough precision for addressing more complicated questions. Therefore, we examined several SRT datasets from different platforms to confirm that SRT data contain meaningful information on CCC.

We first examined a MERSCOPE dataset from a patient with squamous cell carcinoma in the lung. We performed cell typing for this dataset (**Fig. 1a** and **Sup. Fig. 1**) and visualized the distribution of the single cells in their spatial context (**Fig. 1b**). We examined the interaction between fibroblast cells and tumor cells in the EMT pathway. In this pathway, TGFB proteins secreted by cells in the tumor microenvironment are known to induce EMT in tumor cells, with *FN1* and *SNAI2* being important markers of EMT (*14–16*). Then, we enumerated all tumor cell-fibroblast cell pairs (**Fig. 1c** and **Sup. Fig. 2**), and defined cell pairs as "adjacent" if the two cells are within <30μm radius away from each other, and "distant" otherwise. We examined the expression of *FN1* (**Fig. 1d**) and *SNAI2* (**Fig. 1e**) in tumor cells as a function of the averaged expression of *TGFB1*, *TGFB2*, and *TGFB3* in the "adjacent" or "distant" fibroblasts. As expected, **Fig. 1d** and **Fig. 1e** show that the expression of *FN1* and *SNAI2* are positively correlated with the *TGFB1-3* expression in neighboring fibroblasts, but much less so for distant fibroblasts. This suggests that the MERSCOPE platform is capable of supporting the mapping of CCC.

Next, we examined a breast cancer Visium dataset. Unfortunately, unlike the newer MERSCOPE technology, Visium does not capture gene expression at the single-cell resolution. Therefore, we performed clustering analyses and segregated all Visium sequencing spots into a tumor cluster and a broad non-tumor cluster (stromal/immune cells). In **Fig. 1f**, we examined how the expression of tumor cell *CD274* varies as a function of the *PDCD1* pressure from nearby (defined by red arrows) or distant stromal/immune cells. Tumor cells are expected to up-regulate *CD274*, whose protein product PDL1 antagonizes cytotoxic T cells *via* binding to PD1. **Fig. 1g** indeed shows that tumor cells up-regulate *CD274* as a result of higher *PDCD1* expression in neighboring stromal/immune regions, while the same does not hold for PDCD1 pressure from distant stromal/immune cells.

Beyond evaluating SRTs, we also assessed the potential of spatial proteomics for CCC inference. We examined a Cyclic Immunofluorescence (CyCIF) dataset of lymph nodes with metastasis from a human lung adenocarcinoma (*17*) (**Fig. 1h**). In the lymph node, T_h cells aid the maturation of B cells through a direct-contact process called T-B reciprocity (*18–20*). Therefore, for each B cell, we counted the number of T_h cells that are in the B cell's vicinity *vs* T_h cells that are not (adjacency distance cutoff = 100μm). Interestingly, we observed that when there is a larger number of T_h cells in the B cells' neighborhood, the B cells tend to have higher expression

of *MKI67*, a classic proliferation marker, and *CD20*, a B cell maturity marker(*21*). On the other hand, the trend is less clear in cases where T_h cells are distant from the B cells.

Overall, these pieces of evidence suggest that, at least for some of the modern spatial transcriptomics/proteomics technologies, the data quality affords the inference of CCC.

Multi-Instance Learning for Mapping CCCs from SRT Data

Next, we present a Bayesian multiple-instance learning (MIL) model, spacia, to infer CCCs from SRT data. The technical details of spacia are described in **Sup. File 1**. At a high level, spacia has two tiers. The first tier identifies signal-sending cells (senders) that are impacting signal-receiving cells (receivers) based on spatial closeness, indicated by a δ variable (**Fig. 2ab**). As different CCCs could have varying effective spatial ranges, depending on the cell types and types of interactions (contact-based or secretion-based), the model estimates a variable, indicated by b, to allow flexibility in determining δ based on spatial distances. When b is negative, the senders and receivers are determined to have stronger interactions when they are physically closer. In the second tier, spacia discovers gene co-expression patterns between the senders and the receivers (**Fig. 2ab**), but only for senders that are determined to be impacting each receiver cell ($\delta=1$). We infer a variable β that indicates how the expression of the genes/pathways in the senders impacts the genes/pathways in the receiver cells. This model is solved by Markov Chain Monte Carlo (MCMC), which is an iterative process that generates a distribution for the value of each variable of interest so that we can provide both point estimates and inferences of statistical significance.

Importantly, the sending and receiving genes that can be considered by our model do not have to be ligands and receptors, but rather, all genes captured by SRT can be considered for both the sending and receiving portions. This allows us to avoid the questionable rationale of examining the co-expression of ligands/receptors from pre-defined interaction databases for mapping CCC. It also allows us to model the upstream and downstream regulatory signals that occur during CCC. Furthermore, during CCCs, multiple senders in a neighborhood could confer signals to and impact the same receiver. One unique feature of spacia is that it naturally and explicitly integrates this multiple-to-one relationship through MIL. Under a MIL framework, the receiver cells are modeled as “bags” with labels (expression of receiving genes/pathways), and each bag is a collection of instances (senders) characterized by the instance-level features (expression of genes/pathways and spatial closeness to receivers). Spacia infers, for each receiver cell, a set of sender cell(s) that truly interact with this receiver. This unique model design enables us to infer single cell-to-single cell interactions for each cell captured by SRT, unlike scRNA-seq-based approaches that usually only infer interactions between two clusters of cells.

To validate its efficacy, we tested spacia on simulated data. As we demonstrated in **Fig. 2c**, we simulated two types of cells that are interacting. The blue cells are senders, while the red cells are receivers, but the red cells’ expression was simulated to be regulated by only nearby blue

cells. As expected, spacia only infers red and blue cells that are close to each other to be interacting (**Fig. 2c)**. We showed, in **Fig. 2d**, the rate of correct identification of the truly interacting cell pairs, measured by Area Under the ROC curve (AUROC). With more than 10,000 MCMC iterations, the AUROC achieves >0.95 (spacia's default is 50,000 iterations), meaning an almost perfect detection of interacting cell pairs. We then showed in **Fig. 2e** that the b variables were estimated to be negative, consistent with the simulation assumption that only nearby cells are interacting. We also evaluated spacia's estimations of the β variables, in **Fig. 2f** (distributions of βs across all MCMC iterations) and **Fig. 2g** (point estimates of βs). As is shown, the sender genes that were simulated to be truly interacting with the receiver genes had βs that were significantly different from 0, while the converse was true for non-interacting genes. Finally, the posterior samples of the b and β variables demonstrate stable convergence and minimum auto-correlations (**Sup. Fig. 3**). Additional analyses were provided in **Sup. File 2**. Taken as a whole, these results indicate that spacia exhibits excellent statistical properties.

Spacia Validates in SRT Data and Overcomes Limitations of Existing Approaches

Next, we investigated a prostate cancer MERSCOPE dataset and used spacia to infer how the non-tumor cells from the tumor microenvironment impact prostate cancer cells (**Fig. 3a**). We visualized the interacting cell pairs in their spatial locations in **Fig. 3b**. Consistent with our expectation, inferred CCCs are all local. In contrast, the CCCs inferred from CellPhoneDB (*7*) and CellChat (*5*) without spatial context (**Sup. Fig. 4**) indicate numerous CCCs across the entire span of the tissue, which is highly unlikely. More importantly, we noticed that CellPhoneDB and CellChat inferred highly similar sets of interacting genes regardless of the types of cells. For example, in **Fig. 3c**, we showed CellPhoneDB's inference results for three very dissimilar cell types (as sending cells), $CD8^+$ T cells, mast cells, and endothelial cells. Among all unique sending-receiving gene pairs that were inferred to be active in at least one of these sending cell types, 54% exists in all three sending cell types, which suggests an alarming lack of specificity. We investigated this systematically by examining how many predicted interacting genes were shared between all sender cell types (**Fig. 3d**). Of all the CCCs inferred by spacia, 92.7% were unique to one or two sender cell types, with no interactions found to be shared by more than four different sender cell types. In striking contrast, 63.6% and 60.9% of the CCCs are shared among three or more cell types according to CellphoneDB and CellChat, respectively.

Additionally, we attempted to benchmark two recently published methods, COMMOT (*22*) and ncem (*23*), for cellular interaction detection from SRT data. Unfortunately, due to the large number of cells in the MERSCOPE datasets, COMMOT could not be run using the original data as it required more than 2 terabytes of memory, which was much more than what is usually available in high performance computing clusters. After sub-sampling down to no more than 5,000 cells for each cell type, we ran COMMOT with default options using the CellphoneDB database and filtered the interactions to the same cell type pairs as above. Although COMMOT appeared to predict spatially localized CCCs (**Sup. Fig. 4**), it had a very high tendency to produce non-specific interactions (**Fig. 3d**). All CCCs predicted by COMMOT were shared by

three or more different cell types, and 78.3% of interactions were shared among all cell types examined, which is impossible. As for ncem, as of this writing, the Github repository of ncem has provided almost no documentation for practical applications of ncem on new datasets; therefore, it was not tested for this study.

We utilized other forms of data to validate spacia. The NCI Patient-Derived Models Repository (PDMR, pdmr.cancer.gov) provides RNA sequencing data of 70 prostate cancer patient-derived xenografts (PDXs). In PDX models, human immune/stromal cells die out quickly and mouse T/B/NK cells are generally non-existent due to the choice of NSG mouse models (*24*). We separated the bulk PDX gene expression into the human tumor cell *vs.* mouse stromal/immune cell components with Disambiguate (*25*). Next, we used CIBERSORTx (*26*) to further demultiplex the mouse stroma/immune cell component into cell type-specific expression. We then evaluated the correlation between human tumor cells' expression and the expression of mouse fibroblasts and myeloid cells (one example gene pair shown in **Fig. 3e**), which are the two most abundant cell types in the murine components, according to CIBERSORTx. Then we cross-referenced these against the pairs of signal-sending genes in senders (fibroblasts and myeloid cells) and downstream target genes in receiver tumor cells that were detected by spacia. Overall, positively correlated gene pairs from the PDX data are indeed more likely to have positive βs in the corresponding interactions inferred by spacia and *vice versa* (**Fig. 3f**). We also leveraged CytoSig (*27*), which is a manually curated database documenting gene expression changes in many different types of cells upon cytokine treatment. This database contains results for several prostate cancer cell lines (details in the method section). We examined whether the directions of expressional regulations in the interactions inferred by spacia are consistent with the direction of gene expression changes in these prostate cancer cell lines upon cytokine treatment. Due to the nature of the CytoSig database, we can only loop over all possible sender cell types (**Fig. 3a**), with the sender genes being the cytokines from CytoSig in each cell type. We observed top concordance for fibroblasts (overall 71%) and B cells (75%), consistent with the known important roles of these two cell types in modulating the tumor microenvironment through secretion of cytokines and other soluble factors (*28*, *29*). In contrast, we observed the lowest concordance for mast cells (34%). For fibroblasts (**Fig. 3g**), the concordance for TGFB1 (90%) is the highest among all cytokines, consistent with the fact that fibroblasts are one of the major sources of TGFB1 from the tumor microenvironment (*30*, *31*).

Finally, we categorized the interacting gene pairs into "contact-based" and "secretion-based" interactions according to whether the sending gene is known to participate in contact-based or secretion-based interactions (*5*). As expected, the *b*s of cell-cell contact interactions are larger in magnitude (more negative) compared with those of secretion-based interactions (**Fig. 3h**). And cell-cell contact interactions demonstrate larger drop in interaction probability over distance (**Fig. 3i**). Overall, our analyses above validate spacia from the perspective of real data.

Spacia Reveals Induction of Prostate Cancer EMT by the Tumor Microenvironment

TGFBs secreted by fibroblasts induce EMT in tumor cells (*32*, *33*). Inspired by our observations above, we further investigated whether the sender genes are enriched in signaling pathways that are associated with the induction and regulation of EMT in tumor cells. We performed Gene Ontology analysis using GOrilla (*34*, *35*) on the sending genes, for each sender cell type and for each of several receiving genes in the tumor cells that are classical markers or drivers of EMT (*36–44*) (**Fig. 4a**). We found that fibroblasts, B cells, endothelial cells and T_h cells possess the highest numbers of enriched pathways (FDR<0.05) directly or closely related to the regulation of EMT (the full GO results for *JAK1* as an example receiver gene are shown in **Sup. Table 1**). In other words, spacia correctly inferred the upstream signaling pathways in several cell types of the tumor microenvironment that regulate the downstream induction of EMT in the tumor cells.

We focused on a number of secretory ligands that are known to induce EMT activation in tumor cells within the sending cell types above; these ligands include *WNT5A*, *WNT3A*, *TGFB1*, *TGFB2*, *FGF2*, *IL6*, *CXCL8*, *HGF*, and *FGF1 (32, 33, 36, 37, 39–41, 43)*. **Fig. 4b** shows the ranked βs of these ligands in each sender cell type, among all 500 genes captured by MERSCOPE, where a smaller rank refers to stronger regulation of the EMT marker/driver genes. Fibroblasts and endothelial cells demonstrate the strongest activation of tumor EMT *via* these ligands, followed by B cells. While the roles of fibroblasts and endothelial cells in inducing EMT are more well established, it is surprising to see that B cells are also inferred by spacia to induce EMT, though to a lesser extent. Indeed, B cells have been reported to shape the mesenchymal identity of ovarian cancers through the transfer of exosome-derived microRNAs (*45*) and certain subsets of B cells are known to secrete cytokines such as TGFB1 (*45*). We also showed more details of the inferred interactions between these ligands and the EMT genes in a network plot (**Fig. 4c**). It is apparent that fibroblasts, endothelial cells and B cells each employ more than one secretory factor to induce tumor cell EMT. We also uncovered EMT-inducing interactions that have not been described before in prostate cancer. For example, while it has been reported that endothelial cells secrete IL-6 and induce EMT in head and neck tumors (*46*) and esophageal carcinoma (*47*), we showed that such a mechanism also exists in prostate cancers (**Fig. 4c**). We next visualized the spatial co-expression patterns of these ligands in sender cells (**Fig. 4d**) and the EMT levels of the tumor cells (**Fig. 4e**, definition of EMT level in **method section**). In particular, we aggregated over these secretary ligands to form an "EMT activation potential" in each sender, weighted by the βs inferred by spacia, and further aggregated all senders of each receiver cell through a weighted average with weights being the probability of the senders being "primary" ($\delta=1$). **Fig. 4d** confirms that the EMT activation patterns have higher correlation with the EMT levels of the tumor for fibroblasts and endothelial cells (Spearman correlation: ρ=0.822, 0.838), but much lower for B cells (ρ=0.415) (**Fig. 4e**, **Sup. Table 2**). It is interesting, however, that fibroblasts and endothelial cells have largely overlapping spatial patterns of EMT activation, while B cells show a distinct pattern with its EMT activation. It appears that the sum of the EMT activation patterns of fibroblasts/endothelial cells and B cells better corresponds to the EMT level of the tumor cells (**Fig. 4e**).

Our prior work showed that EMT is usually part of a lineage plasticity dys-regulation program in prostate cancer cells (*48*). To determine if the observed increase in EMT activity was also associated with gene expression programs involved in lineage plasticity, we studied the spatial distribution of the stem-like, neuro-endocrine, and basal lineage plasticity levels in the tumor cells. **Fig. 4f** shows that the stem-like and neuro-endocrine lineages are largely correlated with the prostate cancer cell EMT levels (ρ=0.666, 0.854) and the EMT activation potential of fibroblasts and endothelial cells (**Sup. Table 2**). In particular, the correlation between fibroblast EMT activation potential and the tumor cell neuro-endocrine lineage is the highest, achieving a Spearman correlation of 0.916.

To validate the interactions inferred by spacia at the patient level, we generated 21 scRNA-seq datasets from a cohort of prostate cancer patients (**Fig. 4g**). We divided the patients into two subsets of EMT high (N=10) and EMT low (N=11), according to the EMT expression in their tumor cells. We first examined the expression of *FGF2* in fibroblasts, endothelial cells, B cells, and several other immune cell types as controls (**Fig. 4h**). **Fig. 4c** indicates that fibroblasts and endothelial cells are the major cell types that secrete FGF2 and induce prostate cancer cell EMT. Consistent with this observation from MERSCOPE, **Fig. 4h** shows that fibroblasts and endothelial cells are the only two cell types that abundantly express *FGF2*, and more importantly, the expression of *FGF2* is higher in fibroblasts (Pval<0.001) and endothelial cells (Pval<0.001) from patients with EMT-high tumors, compared with EMT-low tumors. We again examined all the secretory factors together by calculating the EMT activation potential as in **Fig. 4d**. As shown in **Fig. 4i**, fibroblasts cells overall possess the strongest EMT activation potential. As expected, fibroblasts and endothelial cells show stronger EMT activation potential in the EMT-high patients, compared with the EMT-low patients (Pval<0.001 for both). B cells also demonstrate the same trend, though not achieving statistical significance. Next, we assessed the association between lineage plasticities of tumor cells and the EMT activation potentials of the stromal/immune cells (**Sup. Fig. 5**), as is done for the MERSCOPE dataset. Consistent with **Sup. Table 2**, the most pronounced observation is that fibroblast EMT activation potential was significantly higher in the group associated with enhanced neuro-endocrine lineage (Pval<0.001). The endothelial EMT activation potential was also higher in the stem-like positive group than in the negative group (Pval<0.001), similarly for the basal lineage (Pval<0.001). Overall, the main discoveries regarding the intra-tumor heterogeneity of EMT and lineage plasticity induction in tumor cells identified by spacia can be extrapolated to inter-tumor differences.

Spacia Infers the Impact of PDL1 Signaling on the Tumor Microenvironment

The analyses above investigated signals sent from the tumor microenvironment. In another application, we deployed spacia to characterize the impact of *PDL1* signaling on the tumor microenvironment. It is well known that *PDL1* and *PD1* interact through protein-to-protein binding, but the transcriptional signaling pathways downstream of *PD1* are not completely clear (*49–52*). We studied a breast cancer MERSCOPE dataset and applied spacia to investigate how the *PDL1* expression of tumor cells, endothelial cells, and macrophages leads to downstream

transcriptional regulation in $CD8^+$ T cells, T_h cells, T_{reg} cells, B cells, and macrophages (**Fig. 5a**). We visualized the pathways enriched in the downstream target genes that spacia inferred for different cell type pairs in **Sup. Fig. 6**. We show that the PDL1-PD1 axis is active in almost all possible combinations of sender cell type-receiver cell type pairs, although the strength of interactions is still the strongest for PDL1 signaling from tumor cells. It is also worth noting that, despite the enrichment of genes in pathways typically related to immune responses, there is also an enrichment related to apoptosis.

To validate these observations, we generated a set of GeoMx data from eight treatment-naive breast cancer patients. Each patient was sampled for an average of five Regions of Interest (ROIs), resulting in a total of 40 ROIs with a mixture of tumor and immune cells. We defined the tumor and immune masks in each ROI (**Fig. 5b**, full images in **Sup. File 2**), and extracted mask-specific gene expression for the tumor cell and immune cell components. We deployed BayesianPrism (*53*) to dissect the immune cell gene expression into the gene expression of each immune cell type for each ROI. As **Fig. 5c** shows, B cells (19%) and $CD8^+$ T cells (10%) are most abundant in these ROIs, followed by T_{reg} cells (6%) and T_h cells (6%), with macrophages accounting for the least proportions (4.5%). Next, we investigated genes in these immune cell types whose expression is positively or negatively correlated with the expression of *PDL1* in tumor cells. We then calculated the overlap between these genes and the *PDL1* downstream target genes that spacia inferred, in terms of direction of expression regulation. Importantly, as BayesPrism's authors suggested, BayesPrism is more accurate for cell types that are abundant in the tissue mixture (only B cells and $CD8^+$ T cells achieve optimal accuracy according to their guidelines). Therefore, we focused this validation on the PDL1 target genes in B cells and $CD8^+$ T cells. As **Fig. 5d** shows, there are statistically significant overlaps between GeoMx and spacia/MERSCOPE, for both B cells (Odds Ratio (OR)=9.3, Pval=0.01) and $CD8^+$ T cells (OR=4.2, Pval=0.011). As expected, the overlap is less pronounced for T_h cells (OR=1.87, Pval=0.14) and T_{reg} cells (OR=1.18, Pval=0.92), and there is no enrichment for macrophages (OR=0.21, Pval=0.37). Even though these two sets of PDL1-downstream genes were derived from two different technologies in two different cohorts of patients, the existence of a significant overlap speaks to the validity of spacia's findings.

We showcase the top genes that are in these overlaps. In $CD8^+$ T cells, both the spacia and the GeoMx analyses indicate that *BCL2* is inhibited by *PDL1* from tumor cells (**Fig. 5e**). *BCL2* is a key anti-apoptosis molecule (*54*) and therefore our results suggest that *PDL1* promotes apoptosis in $CD8^+$ T cells. We also found that *PDL1* down-regulates *GATA3* in $CD8^+$ T cells (**Fig. 5e**), which supports the maintenance and proliferation of T cells downstream of TCR and cytokine signaling (*55*). Both of these observations are in alignment with the well-known immuno-suppressive functions of *PDL1* for $CD8^+$ T cells. For B cells, we found that *PDL1* of tumor cells up-regulates *IDO1* in B cells, in both the spacia and GeoMx analyses (**Fig. 5e**). The role of *IDO1* for B cells is less clear so far, but recent reports have linked *IDO1* with immuno-suppressive roles in B cells (*56*). We also found that *PDL1* up-regulates *PDGFRA* in B

cells (**Fig. 5e**). To our knowledge, there has been no literature on how the expression of *PDGFRA* in B cells is correlated with B cell functions. But we examined the breast cancer scRNA-seq data from Bassez *et al (57)*, where data on various immune cell types pre- and post-anti-PD1 treatment were generated. These data show that *PDGFRA* expression in B cells decreased after anti-PD1 treatment (**Sup. Fig. 7**, Pval=0.025), which also suggests an activating role of *PDL1* for *PDGFRA*.

Next, we applied spacia to a set of pan-cancer MERSCOPE datasets, which include breast cancer, colon cancer, melanoma, lung cancer, liver cancer, ovarian cancer, prostate cancer, and uterine cancer (**Fig. 1ab** and **Sup. File 2**). Here, we still focused on tumor cell-to-$CD8^+$ T cell interactions. Spacia yields a set of downstream target genes for tumor-to-$CD8^+$ T cell interactions in each cancer type, which we term the CD8-PDL1 signatures (**Sup. Table 3**). We first validated these gene signatures, by leveraging the RNA-seq data from The Cancer Genome Atlas Program (TCGA), for the same eight cancer types. We used BayesPrism (*53*) to dissect the TCGA RNA-seq data into cell type-specific gene expression in each patient sample. We investigated the differential gene expression in the $CD8^+$ T cell components in the TCGA patients, between patients with high and low *PDL1* expression in their tumor cell components. Gene Set Enrichment Analysis (GSEA) (*58*) confirmed that the CD8-PDL1 gene signatures identified by spacia are indeed enriched in the top differentially expressed genes. **Fig. 5f** showcases the GSEA results for breast cancer (Pval=0.0007) and ovarian cancer (Pval=0.025). We also calculated a composite score as the weighted average of the genes in each CD8-PDL1 signature, with weights being the inferred βs by spacia, to reflect the overall impact of *PDL1* signaling on $CD8^+$ T cells. For almost all cancer types, patients with higher *PDL1* expression in the tumor cells tend to have higher expression of the CD8-PDL1 signatures (**Fig. 5g**, Pval(liver cancer)=0.012, Pval(lung cancer)=0.69, and P values for all other cancer types <0.002). Overall, these results validate the CD8-PDL1 signatures inferred by spacia.

The PDL1 Signature in $CD8^+$ T Cells Is Prognostic and Predictive

We evaluated the prognostic and predictive powers of the CD8-PDL1 signatures to determine whether they possess any translational value. These signatures can potentially reflect the actual impact of tumor cells' *PDL1* signaling on T cells, which is a more direct measurement of *PDL1* signaling effectiveness and could reveal stronger biological signals compared with testing tumor *PDL1* expression alone.

First, we tested the association between bulk tumor *PDL1* expression and patient overall survival for the eight cancer types of interest in the TCGA patients (**Sup. Fig. 8**). Somewhat counterintuitively, we observed that higher *PDL1* expression tends to predict better survival in these patients (Pval=0.016, all patients combined). One might expect higher *PDL1* expression leads to worse survival as it is known to suppress anti-tumor immune functions. However, inflammatory T-cell responses could incite tumor metastasis (*59*), thus PDL1 can alleviate the risk for metastasis through the inhibition of inflammation. While the underlying biological

mechanisms of this effect are not within the scope of this work, we assessed whether the CD8-PDL1 signatures in $CD8^+$ T cells better capture this phenomenon.

We performed Kaplan-Meier analyses for the CD8-PDL1 signatures in $CD8^+$ T cells in all cancer types combined (**Fig. 6a**), and also for *PDL1* expression in the tumor samples as control (**Fig. 6b**). Here, we also subset the patients into patients with high and low $CD8^+$ T cell infiltrations, based on the $CD8^+$ T cell proportions from BayesPrism. Inspired by our prior observations (*60*, *61*), we posit that the correlation between CD8-PDL1/PDL1 and survival should be stronger when the tumors have more infiltrating $CD8^+$ T cells (otherwise, *PDL1* signaling is irrelevant as there are few T cells to inhibit). In other words, the patient subsets of high *vs.* low $CD8^+$ T cell infiltration serve as another useful control. For all cancer types combined, we found that high CD8-PDL1 signature expression indeed predicts better survival in the $CD8^+$ T cell-high patients (Pval=0.004, OR=0.714, 95% CI=0.564-0.902), whereas there is no significant difference for the $CD8^+$ T cell-low patients (Pval=0.34, OR=0.911, 95% CI=0.750-1.11) (**Fig. 6a**). On the other hand, the expression level of *PDL1* in tumor cells is less prognostic in both patient subsets ($CD8^+$ T cell-high: Pval=0.26, OR=0.862, 95% CI=0.667-1.11; $CD8^+$ T cell-low: Pval=0.16, OR=0.866, 95% CI=0.710-1.05) (**Fig. 6b**). We next limited our analyses to breast cancer only, and again confirmed that the CD8-PDL1 signature is a more robust biomarker than tumor *PDL1* expression for depicting the effect of *PDL1* signaling on patient survival (**Fig. 6c**: CD8-PDL1 signature, CD8+ T cell-high: Pval=0.0002, OR=0.337, 95% CI=0.190-0.597, CD8+ T cell-low: Pval=0.08, OR=0.627, 95% CI=0.371-1.06; **Fig. 6d**: tumor *PDL1* expression, CD8+ T cell-high: Pval=0.46, OR=1.35, 95% CI=0.616-2.94, CD8+ T cell-low: Pval=0.06, OR=0.593, 95% CI=0.347-1.01). We performed multivariate CoxPH analyses to adjust for the effect of confounding clinical covariates. In the pan-cancer cohort, the CD8-PDL1 signature is still significantly predictive of survival in the high $CD8^+$ T cell patients after adjusting for the covariate of different cancer types (**Fig. 6e**, Pval = 0.041, Hazard Ratio=1.28, 95% CI=1.01-1.63). In breast cancer, the CD8-PDL1 signature is also still significantly prognostic in the high $CD8^+$ T cell patients after adjusting for cancer stage (**Fig. 6f**, Pval < 0.0001, Hazard Ratio=2.93, 95% CI=1.6-5.5). The predictiveness (reflected by Hazard Ratios) of CD8-PDL1 signatures diminished in the low $CD8^+$ T cell patients, for either all patients combined (Pval=0.15, Hazard Ratio=1.15, 95% CI=0.95-1.41) or in breast cancer only (Pval=0.06, Hazard Ratio=1.7, 95% CI=0.99-2.9).

Next, we also investigated patients who were treated with anti-PD1/PDL1 therapies. We studied two anti-PD1/PDL1-treated cohorts, Sade-Feldman *et al (62)*, which consists of 32 scRNA-seq datasets generated from peripheral blood of melanoma patients on anti-PD1 or anti-CTLA4+PD1 treatment, and Zhang *et al (63)*, which consists of 11 scRNA-seq datasets generated from various tissue biopsies of breast cancer patients on anti-PDL1 treatment. Both cohorts contained responders and non-responders (as defined in the original works), and the biopsies were collected before and after treatment for scRNA-seq. We performed cell typing for the scRNA-seq data. For the $CD8^+$ T cells in particular, we further classified them into exhausted, effector, effector memory, central memory, and naive T cells, according to marker genes from Sun *et al (64)* in

order to conduct more fine-grained analyses. We observed that exhausted T cells, effector memory T cells, and effector T cells have higher overall expression levels of the CD8-PDL1 signature, with exhausted T cells having the highest expression (**Fig. 6g**). This is expected as the CD8-PDL1 signature measures the immuno-supppressive effect of PDL1 signaling in T cells. In responders, we observed that the CD8-PDL1 signature significantly decreased after treatment in exhausted (Pval=0.005, testing both cohort together) and effector memory T cells (Pval=0.048), while these changes were not observed in non-responders (Pval=0.23 and 0.61 respectively). This contrast indicates that the CD8-PDL1 signature is a candidate biomarker for measuring the effectiveness of immunotherapies that block the PD1/PDL1 axis. Moreover, the melanoma scRNA-seq datasets were generated from patient peripheral blood, which suggests that this biomarker could be measured non-invasively. On the other hand, in the central memory T cells, effector T cells, and naive T cells, the decreases in the CD8-PDL1 signature after treatment are more modest in responders (**Fig. 6g**). Therefore, our analyses indicate that anti-PD1/PDL1 therapies mainly impact exhausted T cells and effector memory T cells.

Finally, we studied the breast cancer scRNA-seq data from Bassez *et al (57)*. In this study, 29 breast cancer patients were treated with pembrolizumab ~9 days before surgical resection of tumors. Paired pre- *vs.* on-treatment biopsies were subjected to scRNA-seq. In this cohort, we were only able to reliably segregate $CD8^+$ T cells into exhausted T cells and other T cells. Again, the exhausted $CD8^+$ T cells showed the highest expression of the CD8-PDL1 signature (**Sup. Fig. 9**). And we observed a decrease, though non-significant, in the CD8-PDL1 signature after treatment. Taken together, the CD8-PDL1 signatures that we defined from SRT data using spacia possess both prognostic and predictive values, which demonstrates the power of these new technologies for yielding novel insights of translational value.

DISCUSSION

We developed spacia to fulfill the unmet gap in detection of cell-to-cell and gene-to-gene interactions from SRT data. scRNA-seq-based mapping of CCC loses single-cell resolution and suffers from high false positive rates, while SRTs provide rich information on gene expression and cell locations to overcome the intrinsic limitations of scRNA-seq. Importantly, we showed that the quality of modern SRT technologies already enables the detection of CCCs with appropriate statistical models. With the increase in throughput and data quality of SRTs, our concept of integrating the transcriptomics and spatial modalities will inspire more and more sophisticated analyses to address exciting scientific questions using SRT data. Take the research in pseudotime and cellular trajectory inference for example. As reviewed and benchmarked by Saelens *et al (65)*, current scRNA-seq-based pseudotime inference tools have very unsatisfactory performances. However, cells in solid tissues grow to form spatially continuous patterns over time. Consideration of the adjacencies of two cells both in their physical and transcriptomic space will likely result in a more informative construction of pseudotime/cellular trajectories.

Whereas most existing CCC inference tools work by examining data against a known database,

spacia focuses on the more general concept of CCC through accurate *in situ* modeling. Spacia achieves this through organically integrating the rich spatial and transcriptional information, which allows it to minimize the number of assumptions and arbitrary parameters. While this approach does not rely on existing knowledge, it offers an unbiased, independent interrogation based on first principles. Furthermore, it enables discoveries that complete known CCC pathways and uncover new pathways in a cellular context specific manner. It is important to note that the target genes identified by spacia, through modeling of transcriptional regulation as a function of sending gene expression, could be either direct targets that are within the core of the interacting pathways or represent more downstream and secondary effects. This could be an advantage or disadvantage depending on one's vantage point. At this time, cross-referencing to existing databases could be helpful for the interpretation of the identified CCCs and distinguishing between "direct target" *vs.* "indirect target", and allows for further interrogation of the regulatory relationships between genes in existing pathways. Therefore, we propose a paradigm of unbiased search of CCCs followed by filtering through prior knowledge in spacia, in contrast to those of prior works, such as CellPhoneDB, whose performance is limited to prior knowledge from the very beginning.

As of this writing, we noticed a few very recently published software for CCC detection from SRT data, such as COMMOT (*22*) and ncem (*23*). However, as we pointed out, it is difficult to execute these software due to technical reasons. Moreover, these software also suffer from other conceptual challenges or deficiencies. For example, COMMOT is still limited to modeling ligand-receptor relationships, which are protein-to-protein interactions. Again, SRT data capture transcriptional changes and the receptor may not necessarily be expressionally regulated by the ligand. COMMOT is also limited to relying its inference on prior databases such as CellPhoneDB in the very beginning of the analyses. On the other hand, ncem cannot explicitly model the strength of regulation from the sender cells/genes to the receiver cells/genes, which is provided by spacia (βs), and it appears that ncem still focuses on cell type-level interactions, rather than inferring single cell-level interactions.

The core of spacia is a fully integrated Bayesian MIL model. Due to the complexity of learning the multiple-to-one functions, MIL problems are much more difficult than typical machine learning problems. Spacia's MIL model allows solving this difficult mathematical problem in a graceful manner. Whereas existing MIL methods mainly focus on predictive performance (*66–68*), our two-tier MIL approach enables concurrent identification of primary instances (the sender cells responsible for the reaction in the receiver cell) and elucidation of the relationships between bags and instances (βs). This technical innovation is important for both the fields of MIL in general and the specific application of finding CCC in SRT data. Critically, in the era of data explosion in biomedical research, MIL can offer elegant solutions to disentangle complex interactions and large, diverse data from various fields. For example, in biochemistry, one might be interested in how the many conformations of a chemical compound are related to its bioactivity (*69*). In real world evidence data, one might want to study how patient risks of

hospitalization can be predicted by the various instances of the patients' prior medical records. We envision that our work will propel the wide adoption of MIL in biomedical research, by providing a success case and also by providing a model that can be further improved upon.

Despite the exciting results presented in our study, mapping CCC is still challenging. Perhaps one of the biggest challenges is associated with the low signal-to-noise ratios (SNRs) of the SRT data. The inaccuracy in cell segmentation is a major contributor to this low SNR. Most, if not all, high-resolution SRT technologies generate the raw expression counts at subcellular or even pixel levels. To aggregate such data to single cell levels, cell boundaries are usually created from the matching H&E staining or fluorescence images *via* segmentation techniques. However, improper cell segmentation can lead to serious errors. For example, two cells that are close by (potentially two cells that are interacting) can be segmented as the same cell. In this scenario, two genes that are interacting with each other from two different cells can be mistaken as a pair of genes that are transcriptionally linked within the same cell. However, the field has seen continuous improvement in the cell segmentation procedure and other pre-processing procedures of SRT data, either in academically generated tools (*70–73*) or commercially available software. We expect such caveats to become less pronounced with future iterations of SRT technologies and the corresponding bioinformatics analysis software.

In summary, we built a general and principled framework for the analysis of CCCs from SRT data that can be applied to a vast number of biological systems. More broadly, when coupled with remarkable experimental and analytical advances in single cell and spatial approaches (*74–83*), spacia will enable us to understand how complex cell states arise from communications in the local cellular community and to move towards holistic models and theories of entire organisms.

METHODS

Simulation data creation

Simulated datasets were generated with the following steps. First, 8,000 cells were generated in a two-dimensional space of 2,000 x 2,000 units. These cells were classified into three types, with the receivers forming the core of several blobs, senders lining the perimeters, and non-interacting cells filling the space between blobs. Senders were divided into two categories: primary senders and non-primary senders based on their distances to a given receiver (distance cutoff = 50). Next, we simulated expression data (50 genes) associated with sender and receiver cells. Expression of each gene of the sender cells was generated with normal distribution (mean = 0, s.d. = 1). The first 5 or 10 genes (in two simulation settings) were designated as truly interacting genes, in two different settings, while the remaining genes were designated as non-interacting genes. Expression for receivers was generated as a weighted sum of gene expression from their primary senders, with weights (βs) generated from uniform distributions. The uniform distribution ranges from 10 to 20 for the truly interacting genes and 0 to 1 for the other genes. The signs of these

weights were randomly assigned to be positive or negative to model upregulation and downregulation of receiving genes by sending genes. Lastly, we added noise to the receiver genes, with noises sampled from a normal distribution (mean = 0, s.d. = 0.5).

MERSCOPE data pre-processing and annotation

The pan-cancer MERSCOPE datasets consisting of 8 different cancer types were downloaded from the publically available MERFISH FFPE Human Immuno-oncology Data Set. The R Seurat package was used to load the "cell_by_gene.csv" file for each MERSCOPE experiment to create a Seurat object for downstream processing and analysis. To ensure we only retain high-quality cells, each experiment was subset so that cells with more than 100 total counts were kept. After the initial clustering and annotation, we subset the T cell population and the epithelial population and performed a second round of the standard workflow for each population to identify the fine-grained clusters for the T cell subpopulations and to segregate epithelial cells into tumor epithelial cells and normal epithelial cells. The final version of the annotated clusters in the UMAP space and the average gene expression of cell type markers in each cluster can be found in **Fig. 1**, **Sup. Fig. 1**, and **Sup. File 2**.

Benchmarking with existing CCC software

CellPhoneDB: Normalized counts and cell type labels were converted into CellPhoneDB compatible TXT files. A microenvironments file was generated to limit the predictions to stromal-tumor interactions only. CellphoneDB 3.1.0 was run using method statistical_analysis, options --counts-data hgnc_symbol, --pvalue 1, and --threads 24 options. Since CellPhoneDB does not designate sending and receiving cells, they were defined by which of the interacting genes was labeled as receptor. For each interaction, the cell type with the gene labeled as receptor was designated as the receiving cell. Interactions where both or neither genes were labeled as receptor were discarded due to ambiguity of the direction of interaction. For consistency with spacia results, results with smooth muscle and normal epithelial cells as sending cells were removed, and only interactions with tumor cells as the receiving cell were kept. Interactions with Pval > 0.05 were removed.

CellChat: CellChat 1.5.0 was used with the default CellChatDB.human ligand-receptor interaction database. Functions *identifyOverExpressedGenes* and *identifyOverExpressedInteractions* were run with default options; *computeCommunProb* was run with type = "truncatedMean" and trim = 0.05; *filterCommunication* used min.cells = 10, and *computeCommunProbPathway*, *aggregateNet*, and *subsetCommunication* were all run with default options. For the outputs, "source" was considered equivalent to spacia's sending cell, "target" equal to receiving cell, "ligand" as sending gene, and "receptor" as receiving gene. For consistency, results with smooth muscle and normal epithelial cells as sending cells were removed, and only interactions with tumor cells as the receiving cell were kept. Interactions with Pval > 0.05 were removed.

Validation with PDMR data

Bulk prostate cancer PDX RNA-sequencing data were downloaded from The NCI Patient-Derived Models Repository (PDMR), NCI-Frederick, Frederick National Laboratory for Cancer Research, Frederick, MD (pdmr.cancer.gov). Disambiguate (*25*) was used to segregate the PDX RNA-seq data into the human and mouse components. CIBERSORTx was used to further deconvolute the mouse stromal/immune component into cell type specific expression values. The CIBERSORTx web portal (cibersortx.stanford.edu) was used and the signature matrix for CIBERSORTx was created on the web portal using the expression matrix and cell typing results from the prostate cancer MERSCOPE data. To comply with the computational limits of the web server, for cell types with large numbers of cells, 5,000 cells were randomly sampled to create the reference sample file. CIBERSORTx was then run in high resolution mode for the PDMR data using default options.

Validation with CytoSig

The CytoSig database was downloaded from cytosig.ccr.cancer.gov. The database was filtered to only include those from prostate cancer cell lines (LNCaP, PC-3M, or MDA-PCa-2b) and involving the following cytokines: *FGF2*, *HGF*, *IL1B*, *IL6*, and *TGFB1*. For cross-referencing spacia results with CytoSig, we assumed the sending cells were each of the sender cell types we tested for spacia (fibroblasts, B cells, *etc*) but the receiving cells are always the prostate cancer cells. This is done due to the nature of the experiments conducted in CyotSig, where receiving cells were treated with soluble cytokine, without sending cells present in the cell culture.

EMT process calculation in the prostate MERSCOPE dataset

We define the "EMT activation potential" scores as the sum of the strengths of EMT-induction signals from nearby sender cells for each tumor cell. This is calculated with the following steps. The β values from the sender-to-tumor cells spacia results are filtered to keep those with sending genes included in a list of secreted factors that have been reported to impact EMT (*HGF*, *WNT3A, WNT5A*, *FGF2*, *IL6*, *CXCL8*, *FGF1*, *TGFB1*, and *TGFB2*), and receiving genes being EMT upstream regulators that have some prior evidence of being impacted by these factors (*JAK1*, *AKT2*, *SMO*, *CTNNB1*, *SMAD2*, and *NFKB2*) (*36–44*). For each sender cell and each sending gene, a set of scores are calculated by multiplying the β of each sending gene-receiving gene pair with the corresponding expression of the sending gene. These scores, one for each receiving gene, are averaged. The averaged scores for all sending genes are further averaged to arrive at a composite score to be assigned as the EMT activation potential of this sender cell. Finally, an EMT activation potential score for each tumor cell is computed by the weighted sum of the EMT activation potentials of the sender cells in each tumor cell bag with the weights being the primary instance probabilities of the senders.

The "EMT score" of a tumor cell is defined to be the mean expression of the EMT marker genes from Gorgola *et al (84)*: *FN1, TWIST1, SNAI1, SNAI2, ZEB1, TGFB1, TGFB2,* and *CTNNB1*,

which represents the activity level of EMT in that tumor cell. We performed additional analyses to validate that this EMT score is valid in our scRNA-seq datasets (**Sup. File 2**).

Lineage score calculation in the prostate cancer MERSCOPE dataset

We collected gene markers from Deng *et al (48)* to calculate the lineage plasticity scores for the basal, neuro-endocrine and stem-like lineages. We were not able to investigate the luminal lineage, as the most important canonical markers such as *AR* and *KLK3* were missing. For the basal lineage, *TP63*, *CAV1*, and *LAMB3* were used. For the neuro-endocrine lineage, *EZH2* and *NCAM1* were used. For the stem-like lineage, *KIT*, *LGR5*, and *LGR6* were used.

Prostate patients for scRNA-seq

The study was performed following protocols approved by the Institutional Review Board of the University of Texas Southwestern Medical Center. There are two studies from which we obtained patient biopsies. STU 072010-098: Tissue Procurement and Outcome Collection for Radiotherapy Treated Patients & Healthy Participants, and STU062014-027: Phase I Clinical Trial of Stereotactic Ablative Radiotherapy (SABR) of Pelvis and Prostate Targets for Patients with High Risk Prostate Cancer. We obtained a total of 21 androgen deprivation (ADT)-treated patients, 4 untreated prostate cancer patients, and 4 healthy donors. The scRNA-seq experiments were done in Dr. Douglas Strand's lab following the protocol in Henry *et al (85)*. Briefly, we performed a 1 hour digestion with 5mg/ml collagenase type I, 10mM ROCK inhibitor, and 1mg DNase. 3' GEX barcoding was performed on a 10X controller and sequencing was performed on an Illumina NextSeq 500 sequencer. These 29 scRNA-seq datasets were processed through QC, integration, and cluster annotation prior to the analysis. We utilized the R DoubletFinder package and followed the recommended workflow for filtering the doublets or multiplets in the data. For the joint analysis of multiple patient scRNA-seq datasets, we followed the Seurat integration vignette. Each dataset was processed through the *ScaleData* and *RunPCA* functions. The anchors for integration were found with the "rpca" mode of the *FindIntegrationAnchors* function. Utilizing these anchors, we proceeded with the *IntegrateData* function for the integration of the 29 scRNA datasets. Initially, the clusters were automatically annotated with the R SingleR package and the annotated dataset from Song *et al* (*86*).

(1) The patients were dichotomized into EMT+ and EMT- groups with the following steps. First, the EMT level of each tumor cell is defined as above in the prostate cancer MERSCOPE dataset. The global median EMT level is calculated from all the tumor cells. Then, for each patient, the percentage of tumor cells with higher EMT levels than the global median cutoff is calculated. The patients were divided into a EMT+ group and a EMT- group based on the median of these percentages. (2) The calculation of the EMT activation potential in the scRNA-seq data follows the same method as in the MERSCOPE dataset, except for that we calculated the activation potential for each sender cell, and did not aggregate to receiver cells through spatial averaging (as these are not SRT data). (3) Finally, the lineage scores for basal, neuro-endocrine, and

stem-like were calculated as in the MERSCOPE dataset, utilizing the complete marker set from Deng *et al (48)*.

Breast cancer patients for GeoMx

The study was performed following protocols approved by the Institutional Review Board of the University of Texas Southwestern Medical Center. The IRB Protocol Number is STU2018-0015: Pre-surgical trial of letrozole in post-menopausal patients with operable hormone-sensitive breast cancer. Formalin-fixed, paraffin-embedded (FFPE) tumor tissues were obtained from diagnostic core needle biopsies from postmenopausal patients with stage I to stage III operable ER+/HER2- breast cancer enrolled in a clinical trial (UT Southwestern SCCC-11118).

To perform the GeoMx® assays, 5 µm FFPE sections were mounted on charged slides, baked, and prepared on the Leica Biosystem, following the manufacturer's automated slide preparation user manual. After hybridization with RNA probes that are conjugated to barcoded oligonucleotide tags with an ultraviolet (UV) photocleavable linker and staining with fluorescent morphology markers consisting of pan-cytokeratin (epithelial and tumoral regions), CD45 (immune cells), SYTO13 (nuclear) and Ki67 (proliferation marker), the slides were loaded onto a GeoMx® digital spatial profiler (DSP, NanoString Technologies) and scanned. After labeling, the tissues were imaged, and we selected specific regions of interest (ROIs) sized 222 x 354.6 µm after consultation with a pathologist. Each ROI was further segmented into Areas of Illumination (AOI) based on morphological features. Subsequently, the selected areas were exposed to UV light, and the barcoded oligos were released, aspirated, and dispensed into a collection plate for library construction for next-generation sequencing (NGS).

GeoMx NGS libraries were prepared according to the manufacturer’s instructions. Briefly, after the collection of the probes was completed, aspirates in the collection plate were dried at 65°C for 1 hour in a thermal cycler with an open lid and resuspended in 10 µL of nuclease-free water. 4 µL of rehydrated aspirates were mixed with 2 µL of 5×PCR Master Mix and 4 µL of SeqCode primers. PCR amplification was then performed with 18 cycles. The indexed libraries were pooled equally and purified twice with 1.2×AMPure XP beads (Beckman Coulter). The final libraries were evaluated and quantified using Agilent's High Sensitivity DNA Kit and Invitrogen's Qubit dsDNA HS assay, respectively. Total sequencing reads per DSP collection plate were calculated using the NanoString DSP Worksheet. The libraries were sequenced using 38 bp paired-end sequencing (PE 38) on an Illumina NovaSeq 6000 system with a 100-cycle S1 kit (v1.5). FASTQ files were processed into digital count conversion digital files using Nanostring’s GeoMx NGS Pipeline software. Quality control, data filtering and normalization (Q3) were performed using the GeoMx DSP Data Analysis suite.

BayesPrism bulk expression data deconvolution

The BayesPrism R package was installed according to its authors’ instructions at

github.com/Danko-Lab/BayesPrism. The histology-matching MERSCOPE dataset was used as the single cell reference in order to maximize consistency with spacia's results. BayesPrism was run using the *run.prism* function with default options. The cell type assignments we produced for the MERSCOPE datasets were used as the "cell.type.labels". Cell type fractions were extracted using the *get.fraction* function. Predicted cell type-specific expression was extracted by running the *get.exp* function on each cell type.

The Cancer Genome Atlas Program (TCGA) RNA-seq data analysis

TCGA data of eight cancer types that are of the same histologies as the MERSCOPE data were downloaded from Broad GDAC Firehose. These include: BRCA, COAD, LIHC, LUSC, OV, PRAD, SKCM, and UCEC. We only considered primary tumor samples of the TCGA cohort since all MERSCOPE datasets are from primary tumor samples. As we expected, inclusion of the TCGA metastatic samples (data not shown) in all downstream analyses yields similar results but with less statistical strength. BayesPrism was then run with default options, using the corresponding MERSCOPE data as the reference.

To define the CD8-PDL1 signature for each cancer type, we extracted, from spacia's results, all receiving genes with $b < 0$, b Pval < 0.01, and β Pval <0.1. The genes of each cancer type were further filtered to keep those that have appeared in the gene list from breast cancer and at least two other cancer types. This was done due to our promising analysis results in breast cancer in **Fig. 5a-e**. The genes that passed all filters were designated as the CD8-PDL1 signature for each cancer type. The R Fast Gene Set Enrichment Analysis (fgsea) package was used to perform GSEA using the CD8-PDL1 signatures for the TCGA BayesPrism-dissected expression data. GSEA was run using the *fgsea* function with options eps=0, minsize=10, and scoreType="pos". The results were plotted using the *plotEnrichmentData* function from the fgsea package and the R ggplot2 package.

TCGA patient survival analysis

The same TCGA datasets and CD8-PDL1 signature genes as above were used. For each patient sample in the TCGA datasets, the CD8-PDL1 signature expression level was calculated as the dot product of the normalized and log-transformed gene expression values of the BayesPrism-dissected $CD8^+$ T cells and the β values of the genes in the corresponding CD8-PDL1 signature gene set. Survival analyses were performed using the *survfit* function from the R survival package with the CD8-PDL1 signature or tumor *PDL1* as the covariate. The Kaplan-Meier curves were plotted using the *ggsurvplot* function from the R *survminer* package. For the forest plots, the *coxph* function from the R *survival* package was used to fit Cox proportional hazards regression models with CD8-PDL1 signature and cancer type or stage as covariates. The forest plots were then generated using the *ggforest* function from the *survminer* package.

Anti-PD1/PDL1 scRNA-seq data analyses

The Zhang *et al (63)* study originally identified naive (CD8Tn), effector (CD8Teff), and effector memory (CD8Tem) $CD8^+$ T cell subsets in the patient peripheral blood, while the Sade-Feldman *et al (62)* study originally classified CD8Tem, central memory (CD8Tcm), exhausted (CD8Tex) subtypes. To homogenize classification and nomenclature, we loaded each dataset as a Seurat object to recluster and tried to identify all $CD8^+$ T subtypes in both studies. Apart from the 20 dimensions used in FindNeighbors and FindClusters, all default parameters were used. Utilizing the gene marker sets from Sun *et al (64)*, we annotated 28 clusters in the melanoma dataset and 13 clusters in the breast cancer dataset, compared to 4 and 6 clusters originally. All $CD8^+$ T subtypes were identified, except for CD8Tcm in the Zhang *et al* dataset. The CD8-PDL1 signature score was calculated as in the TCGA data analysis, except that we have the actual $CD8^+$ T cell gene expression in these scRNA-seq data, as opposed to BayesPrism-dissected expression for the TCGA samples.

Statistics & reproducibility

Computations were performed in the R (3.6.3 and 4.1.3) and Python (3.7 and 3.8) programming languages. All statistical tests were two-sided unless otherwise described. Spacia was run in the PCA mode (see our documentation) for all our analyses in this work. Unless otherwise stated, the spacia results were filtered to keep those interactions satisfying $b < 0$, b Pval<1x10E-30, and β Pval<0.01. For pre-processing of all SRT and scRNA-seq data, we loaded the original data into Seurat objects and analyzed them through a standard workflow, using *NormalizeData*, *FindVariableFeatures*, *ScaleData*, *RunPCA*, *FindNeighbors*, *FindClusters*, and *RunUMAP* functions, before other custom operations. We executed the pipeline with default parameters except for dims = 1:20 for *FindNeighbors* and *RunUMAP* in the prostate cancer scRNA-seq datasets to capture sufficient variability in the principal components. GOrilla (*34*, *35*) pathway enrichment analyses were performed by inputting the genes on the MERSCOPE gene panel into GOrilla, where the genes were filtered to keep only those that satisfy $b < 0$ and sorted by β values for each sending cell type (tumor cells are the receiving cells). The Kernel Density Estimation for the visualization of spatial EMT patterns was calculated from the *gaussian_kde* function from the SciPy package with bandwidth less than the silverman factor, to demonstrate robustness from the spatial noise.

Data and code availability statement

The MERSCOPE datasets were downloaded from vizgen.com/data-release-program. The breast cancer Visium dataset was downloaded from www.10xgenomics.com/resources/datasets. The CyCIF dataset was downloaded from www.synapse.org/#!Synapse:syn19003074. The PDX RNA-seq datasets were downloaded from pdmdb.cancer.gov/web/apex/f?p=101:41:0:. The CytoSig data were downloaded from cytosig.ccr.cancer.gov. The TCGA data were downloaded from gdac.broadinstitute.org. The scRNA-Seq datasets by Zhang *et al (63)* and Sade-Feldman *et al (62)* were accessed *via* Gene Expression Omnibus with accession numbers GSE169246 and

GSE120575, respectively. The scRNA-Seq datasets by Bassez *et al(57)* were accessed from biokey.lambrechtslab.org.

The prostate cancer scRNA-seq data that we generated were archived at doi.org/10.5281/zenodo.8270765. The breast cancer GeoMx data that we generated were archived at github.com/yunguan-wang/Spacia/data. The spacia software is available at the Database for Actionable Immunology (*18*, *82*, *87*) (dbai.biohpc.swmed.edu, Tools page).

FUNDING

This study was supported by the National Institutes of Health (NIH) [R01CA258584/TW, RC2DK129994/DS, R01DK115477/DS, R01DK135535/DS, R01CA222405/SZ, R01CA255064/SZ], Cancer Prevention Research Institute of Texas [RP230363/TW, RP190208/TW, RR170061/CA,RR220024/SZ], Dedman Family Scholars in Clinical Care [ND].

ACKNOWLEDGMENTS

We acknowledge Dr. Shijia Zhu for creating **Fig. 2c**.

AUTHOR CONTRIBUTION STATEMENT

J.Z., Y.W., W.C. contributed to all bioinformatics analyses and implemented the software. A.M., J.G., R.H., P.M., D.S., N.D. generated the prostate cancer scRNA-seq data and/or provided critical insights in analyses. M.Z., F.N., L.G., N.U., A.H., C.A, generated the breast cancer GeoMX datasets. M.Z., S.Z., Z.Z., D.C. performed *in situ* ISS analyses on breast cancer mouse models. F.W. created the readthedocs website. G.X., W.X., Y.X., T.W. developed the initial concept and provided resources for the study. All authors wrote the manuscript.

COMPETING INTEREST STATEMENT

Tao Wang is one of the scientific co-founders of NightStar Biotechnologies, Inc. Ariella Hanker receives or has received research grants from Takeda and Lilly and nonfinancial support from Puma Biotechnology and Tempus. Carlos Arteaga receives or has received research grants from Pfizer, Lilly, and Takeda; holds minor stock options in Provista; serves or has served in an advisory role to Novartis, Merck, Lilly, Daiichi Sankyo, Taiho Oncology, OrigiMed, Puma Biotechnology, Immunomedics, AstraZeneca, Arvinas, and Sanofi; and reports scientific advisory board remuneration from the Susan G. Komen Foundation.

FIGURE AND TABLE LEGENDS

Fig. 1 Rationale for spacia. (a) A heatmap depicting the cell typing results of the lung cancer MERSCOPE dataset. Tumor subtype #1 accounts for most of the tumor cells. (b) The distribution of the different types of cells within their spatial context. (c) A zoomed-in view of one region of the lung cancer MERSCOPE dataset. Potentially interacting tumor-fibroblast cell pairs, defined based on spatial distances, are connected by green

lines. (d,e) The expression levels of *FN1* (d) and *SNAI2* (e) in tumor cells as a function of *TGFB* expression in adjacent or distant fibroblasts. Adjacent fibroblasts are the connected fibroblasts in panel (c), and *vice versa*. (f) The breast cancer Visium dataset. Green color denotes the stromal/immune cell spots and the orange color denotes the tumor cell spots. Potentially interacting tumor-stromal/immune spots, defined based on spatial distances, are connected by red arrows. (g) Tumor *CD274* expression as a function of stromal *PDCD1* expression in adjacent Visium sequencing spots or distant spots. Adjacent spots were defined as in (f). (h) Visualization of the lymph node CyCIF dataset. (i) B cell Ki67 and CD20 protein expression as a function of the number of adjacent or distant T_h cells.

Fig. 2 The spacia model. (a) Cartoons explaining the concept of "primary instances", namely the sending cells that are truly interacting with the receiver cells. The purple senders refer to senders that interact with receivers through cell-to-cell contact. The green senders refer to senders that interact with receivers through secreted ligands. (b) Diagram showing the key elements of the model structure of spacia. (c) Inference results of spacia on a simulation dataset. Blue color refers to sender cells, red color refers to receiver cells and green arrows refer to CCC. (d) ROC curves measuring the accuracy of spacia finding the correct primary instances in the simulation data, with increasing numbers of total MCMC chains. (e) The distribution of the b variables across MCMC iterations after stabilization, in two MCMC chains. (f) The distribution of the β variables across MCMC iterations after stabilization, for genes in senders that are (top) or are not (bottom) truly interacting with receiving genes. (g) Point estimates of the β variables. Left: 5 truly interacting genes in the simulation data; right: 10 truly interacting genes.

Fig. 3 Validating spacia in real data. (a) Spacia inferred CCCs from select immune and stromal cells to tumor cells in the prostate cancer MERSCOPE dataset. The spacia results were filtered to only visualize those that satisfy $b < 0$, b Pval<1x10E-30, β Pval<0.001, and top 1% of z-scored β across all results. (b) Spatial representation of the spacia results. Cells are color labeled by cell types, and CCCs are indicated in black. (c) Overlap between the predicted CCCs by CellphoneDB on the same MERSCOPE dataset for three example sending cell types. The numbers indicate the proportions of total CCCs found in each overlap. Some example CCCs that are inferred to exist in all three cell types are listed. (d) Comparison of the degree of overlap in inferred CCCs between different sending cell types, by spacia, CellChat, CellphoneDB, and COMMOT. The Y axis refers to the proportions of interactions shared in n cell types *vs.* all predicted interactions. (e) An example scatterplot showing the correlations between the expression of sending and receiving genes in their respective cell types in the PDX RNA-seq data. The dotted lines indicate 95% CI. (f) Spacia β values are more likely to be positive for sending genes and receiving genes that are positively correlated in the PDX data, and *vice versa*. 95% CI is indicated with the dotted lines. (g) Concordance between spacia's inferred fibroblast-to-cancer cell interactions (only considering sign of β) and the direction of

expressional changes of prostate cancer cell lines after cytokine treatment, recorded by the CytoSig database. (h) Difference in the spatial range of interactions (magnitude of *b*) between spacia-inferred CCCs that are secretion-based or contact-based. (i) Probability of interaction as a function of distance from the sending cells to the receiving cell, for secretion-based CCCs (left) and contact-based CCCs (right).

Fig. 4 Applying spacia to reveal EMT and lineage plasticity induction signals from the prostate cancer tumor microenvironment. (a) Sankey diagram describing GO term enrichment in the sending genes, of each sender cell type, that are inferred by spacia to be impacting well known EMT marker genes in the tumor receiver cells. The width of the flow is scaled with the P values, and the parent term is connected with the child term if both terms were identified between the same sender cell-receiving gene pair. (b) The rankings of βs for known cytokine ligands that could induce EMT, among all the sending genes input into spacia, for each cell type. A smaller rank refers to a stronger interaction strength (larger β). (c) The top sending-receiving gene pairs inferred by spacia, for fibroblasts, endothelial cells and B cells as sending cell types. (d) The kernel density estimation of the EMT activation potentials in fibroblasts, endothelial cells and B cells. (e) The kernel density estimation of the EMT levels in prostate cancer cells. (f) The kernel density estimations of the stem-like, neuro-endocrine, and basal lineage plasticity scores in the prostate cancer cells. (g) UMAP plot showing the cell types of single cells from our prostate cancer patients. (h) The *FGF* expression of different sending cell types in the EMT+ and EMT- patients. (i) The EMT activation potentials of different sending cell types in the EMT+ and EMT- patients.

Fig. 5 Spacia reveals *PDL1* downstream target genes. (a) Sending and receiving cell type pairs that were analyzed by spacia for the breast cancer MERSCOPE dataset. (b) Left: One example ROI of the GeoMX data; right: the region of the H&E slide corresponding to the same ROI. (c) Abundances of each type of immune cells in the GeoMx data, as predicted by BayesPrism. Cell types are ordered by their average abundance across the 40 ROIs. (d) Odds ratios showing the overlap between spacia-inferred *PDL1* downstream target genes and genes that are differentially expressed in each type of immune cells in the GeoMx data, comparing ROIs that are PDL1+ and PDL1- in the tumor cells. (e) Expression of several representative receiving genes that are in the overlap of (d) for B cells and $CD8^+$ T cells, as a function of tumor cell *PDL1* expression, in the GeoMx data. (f) Gene set enrichment analysis (GSEA) in the TCGA breast and ovarian cancer datasets to evaluate if the corresponding CD8-PDL1 signature genes are indeed among the top genes that are differentially expressed between PDL1+ and PDL1- patients. (g) Expression levels of the CD8-PDL1 signatures in $CD8^+$ T cells in the TCGA samples, dichotomized by tumor cell *PDL1* expression.

Fig. 6 The PDL1-CD8 signature is prognostic and predictive. (a) Prognostic value of the PDL1-CD8 signature for overall survival of TCGA patients of all eight cancer types. (b)

Prognostic value of tumor *PDL1* expression for overall survival of TCGA patients of all eight cancer types. (c) Prognostic value of the PDL1-CD8 signature for overall survival of TCGA breast cancer patients. (d) Prognostic value of tumor *PDL1* expression for overall survival of TCGA breast cancer patients. (e) Forest plot of hazard ratios from a Cox proportional hazard (CoxPH) model considering cancer types as confounding variables for patients of all eight cancer types. (f) Forest plot of hazard ratios from a CoxPH model considering cancer stages as confounding variables for breast cancer patients. (g) The CD8-PDL1 signature expression levels of $CD8^+$ T cells in responders/non-responders and in pre-/post-treatment samples from the Zhang *et al* study and Sade−Feldman *et al* study.

Sup. Fig. 1 UMAP plot showing the distribution of the different types of cells in their gene expression space.

Sup. Fig. 2 Zoomed-in views of four more regions of the lung cancer MERSCOPE dataset. Potentially interacting tumor-fibroblast cell pairs, defined based on spatial distances, are connected by green lines.

Sup. Fig. 3 Traceplots and autocorrelation plots prove convergence and stability of the MCMC estimation process in spacia. Only MCMC iterations after the burn-in period are shown.

Sup. Fig. 4 Visualizing the CCCs predicted by spacia, CellPhoneDB, CellChat, and COMMOT in their spatial context and at the single cell level. To reduce cluttering, for each sender-receiver cell type pair, 10 connections were selected at random and visualized for CellPhoneDB's results, and 500 connections were selected at random and visualized for CellChat's results.

Sup. Fig. 5 EMT activation potentials of each sending cell type by patient groups, dichotomized according to their lineage plasticity levels, for each lineage, in the prostate cancer cells.

Sup. Fig. 6 Overlap of *PDL1* regulated genes with the MsigDB Hallmark pathways. Genes in receiver cells that were significantly regulated by *PDL1* from sender cells were intersected with the genes of each MsigDB Hallmark pathway. The ratio values represent the proportions of all genes of each pathway that are intersecting.

Sup. Fig. 7 The expression of *PDGFRA* in B cells before and after anti-PD1 treatment, in the Bassez cohort.

Sup. Fig. 8 Higher tumor *PDL1* expression is associated with better overall survival in TCGA patients of all eight cancer types, separately or combined. Patients were dichotomized by bulk tumor *PDL1* expression

Sup. Fig. 9 Expression level of the CD8-PDL1 signature in $CD8^+$ T cells of the Bassez cohort, in pre-/post-treatment samples and exhausted/other T cell subsets.

Sup. Table 1 GO ontology analyses to identify the top enriched pathways in the sending genes of each sending cell type with the most significant interactions with *JAK1* as receiver gene in the tumor cells.

Sup. Table 2 Correlations between EMT activation potentials of fibroblasts, endothelial cells and B cells and the EMT levels and lineage plasticity levels of the prostate cancer cells.

Sup. Table 3 CD8-PDL1 signature genes in all eight cancer types.

Sup. File 1 Mathematical and implementation details of spacia.

Sup. File 2 Additional bioinformatics analyses associated with this study.

Bibliography

1. T. J. Bechtel, T. Reyes-Robles, O. O. Fadeyi, R. C. Oslund, Strategies for monitoring cell-cell interactions. *Nat. Chem. Biol.* **17**, 641–652 (2021).
2. E. Armingol, A. Officer, O. Harismendy, N. E. Lewis, Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
3. D. Dimitrov, D. Türei, M. Garrido-Rodriguez, P. L. Burmedi, J. S. Nagai, C. Boys, R. O. Ramirez Flores, H. Kim, B. Szalai, I. G. Costa, A. Valdeolivas, A. Dugourd, J. Saez-Rodriguez, Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).
4. S. Wang, H. Zheng, J. S. Choi, J. K. Lee, X. Li, H. Hu, A systematic evaluation of the computational tools for ligand-receptor-based cell-cell interaction inference. *Brief. Funct. Genomics*. **21**, 339–356 (2022).
5. S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, Q. Nie, Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
6. R. Browaeys, W. Saelens, Y. Saeys, NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods*. **17**, 159–162 (2020).
7. M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
8. R. Hou, E. Denisenko, H. T. Ong, J. A. Ramilowski, A. R. R. Forrest, Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).
9. S. Cabello-Aguilar, M. Alame, F. Kon-Sun-Tack, C. Fau, M. Lacroix, J. Colinge, SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48**, e55 (2020).
10. R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, F. Chen, Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).

11. S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, E. Z. Macosko, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. **363**, 1463–1467 (2019).

12. C.-S. Cho, J. Xi, Y. Si, S.-R. Park, J.-E. Hsu, M. Kim, G. Jun, H. M. Kang, J. H. Lee, Microscopic examination of spatial transcriptome using Seq-Scope. *Cell*. **184**, 3559-3572.e22 (2021).

13. Y. Lee, D. Bogdanoff, Y. Wang, G. C. Hartoularos, J. M. Woo, C. T. Mowery, H. M. Nisonoff, D. S. Lee, Y. Sun, J. Lee, S. Mehdizadeh, J. Cantlon, E. Shifrut, D. N. Ngyuen, T. L. Roth, Y. S. Song, A. Marson, E. D. Chow, C. J. Ye, XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. *Sci. Adv.* **7** (2021), doi:10.1126/sciadv.abg4755.

14. A. Sundqvist, E. Vasilaki, O. Voytyuk, Y. Bai, M. Morikawa, A. Moustakas, K. Miyazono, C.-H. Heldin, P. Ten Dijke, H. van Dam, TGFβ and EGF signaling orchestrates the AP-1- and p63 transcriptional regulation of breast cancer invasiveness. *Oncogene*. **39**, 4436–4449 (2020).

15. Y. Yoshimatsu, T. Watabe, Emerging roles of inflammation-mediated endothelial-mesenchymal transition in health and disease. *Inflamm. Regen.* **42**, 9 (2022).

16. K. Aomatsu, T. Arao, K. Sugioka, K. Matsumoto, D. Tamura, K. Kudo, H. Kaneda, K. Tanaka, Y. Fujita, Y. Shimomura, K. Nishio, TGF-β induces sustained upregulation of SNAI1 and SNAI2 through Smad and non-Smad pathways in a human corneal epithelial cell line. *Invest. Ophthalmol. Vis. Sci.* **52**, 2437–2443 (2011).

17. Z. Du, J.-R. Lin, R. Rashid, Z. Maliga, S. Wang, J. C. Aster, B. Izar, P. K. Sorger, S. Santagata, Qualifying antibodies for image-based immune profiling and multiplexed tissue imaging. *Nat. Protoc.* **14**, 2900–2930 (2019).

18. J. Zhu, A. Gouru, F. Wu, J. A. Berzofsky, Y. Xie, T. Wang, BepiTBR: T-B reciprocity enhances B cell epitope prediction. *iScience*. **25**, 103764 (2022).

19. J. A. Berzofsky, T-B reciprocity. An Ia-restricted epitope-specific circuit regulating T cell-B cell interaction and antibody specificity. *Surv. Immunol. Res.* **2**, 223–229 (1983).

20. J. A. Berzofsky, G. K. Buckenmeyer, G. Hicks, F. R. Gurd, R. J. Feldmann, J. Minna, Topographic antigenic determinants recognized by monoclonal antibodies to sperm whale myoglobin. *J. Biol. Chem.* **257**, 3189–3198 (1982).

21. K. Kläsener, J. Jellusova, G. Andrieux, U. Salzer, C. Böhler, S. N. Steiner, J. B. Albinus, M. Cavallari, B. Süß, R. E. Voll, M. Boerries, B. Wollscheid, M. Reth, CD20 as a gatekeeper of the resting state of human B cells. *Proc Natl Acad Sci USA*. **118** (2021), doi:10.1073/pnas.2021342118.

22. Z. Cang, Y. Zhao, A. A. Almet, A. Stabell, R. Ramos, M. V. Plikus, S. X. Atwood, Q. Nie, Screening cell-cell communication in spatial transcriptomics via collective optimal transport. *Nat. Methods*. **20**, 218–228 (2023).

23. D. S. Fischer, A. C. Schaar, F. J. Theis, Modeling intercellular communication in tissues

using spatial graphs of cells. *Nat. Biotechnol.* **41**, 332–336 (2023).

24. T. Wang, R. Lu, P. Kapur, B. S. Jaiswal, R. Hannan, Z. Zhang, I. Pedrosa, J. J. Luke, H. Zhang, L. D. Goldstein, Q. Yousuf, Y.-F. Gu, T. McKenzie, A. Joyce, M. S. Kim, X. Wang, D. Luo, O. Onabolu, C. Stevens, Z. Xie, J. Brugarolas, An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discov.* **8**, 1142–1155 (2018).

25. M. J. Ahdesmäki, S. R. Gray, J. H. Johnson, Z. Lai, Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. [version 2; peer review: 3 approved]. *F1000Res.* **5**, 2741 (2016).

26. A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

27. P. Jiang, Y. Zhang, B. Ru, Y. Yang, T. Vu, R. Paul, A. Mirza, G. Altan-Bonnet, L. Liu, E. Ruppin, L. Wakefield, K. W. Wucherpfennig, Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. *Nat. Methods*. **18**, 1181–1191 (2021).

28. N. M. de Gruijter, B. Jebson, E. C. Rosser, Cytokine production by human B cells: role in health and autoimmune disease. *Clin. Exp. Immunol.* **210**, 253–262 (2022).

29. S. Davidson, M. Coles, T. Thomas, G. Kollias, B. Ludewig, S. Turley, M. Brenner, C. D. Buckley, Fibroblasts as immune regulators in infection, inflammation and cancer. *Nat. Rev. Immunol.* **21**, 704–717 (2021).

30. K. M. Hargadon, Dysregulation of TGFβ1 Activity in Cancer and Its Influence on the Quality of Anti-Tumor Immunity. *J. Clin. Med.* **5** (2016), doi:10.3390/jcm5090076.

31. X. Shi, J. Yang, S. Deng, H. Xu, D. Wu, Q. Zeng, S. Wang, T. Hu, F. Wu, H. Zhou, TGF-β signaling in the tumor metabolic microenvironment and targeted therapies. *J. Hematol. Oncol.* **15**, 135 (2022).

32. V. Tirino, R. Camerlingo, K. Bifulco, E. Irollo, R. Montella, F. Paino, G. Sessa, M. V. Carriero, N. Normanno, G. Rocco, G. Pirozzi, TGF-β1 exposure induces epithelial to mesenchymal transition both in CSCs and non-CSCs of the A549 cell line, leading to an increase of migration ability in the CD133+ A549 cell fraction. *Cell Death Dis.* **4**, e620 (2013).

33. Q. Ping, C. Wang, X. Cheng, Y. Zhong, R. Yan, M. Yang, Y. Shi, X. Li, X. Li, W. Huang, L. Wang, X. Bi, L. Hu, Y. Yang, Y. Wang, R. Gong, J. Tan, R. Li, H. Li, J. Li, R. Li, TGF-β1 dominates stromal fibroblast-mediated EMT via the FAP/VCAN axis in bladder cancer cells. *J. Transl. Med.* **21**, 475 (2023).

34. E. Eden, R. Navon, I. Steinfeld, D. Lipson, Z. Yakhini, GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*. **10**, 48 (2009).

35. E. Eden, D. Lipson, S. Yogev, Z. Yakhini, Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* **3**, e39 (2007).

36. J. Cai, L. Xia, J. Li, S. Ni, H. Song, X. Wu, Tumor-Associated Macrophages Derived TGF-β–Induced Epithelial to Mesenchymal Transition in Colorectal Cancer Cells through Smad2,3-4/Snail Signaling Pathway. *Cancer Res. Treat.* **51**, 252–266 (2019).

37. D. Sun, T. Luo, P. Dong, N. Zhang, J. Chen, S. Zhang, L. Dong, H. L. A. Janssen, S. Zhang, M2-polarized tumor-associated macrophages promote epithelial-mesenchymal transition via activation of the AKT3/PRAS40 signaling pathway in intrahepatic cholangiocarcinoma. *J. Cell. Biochem.* **121**, 2828–2838 (2020).

38. W. Zhang, J. Gu, J. Chen, P. Zhang, R. Ji, H. Qian, W. Xu, X. Zhang, Interaction with neutrophils promotes gastric cancer cell migration and invasion by inducing epithelial-mesenchymal transition. *Oncol. Rep.* **38**, 2959–2966 (2017).

39. J. Qu, T. Cheng, L. Liu, J. Heng, X. Liu, Z. Sun, W. Wang, K. Li, N. Yang, Mast cells induce epithelial-to-mesenchymal transition and migration in non-small cell lung cancer through IL-8/Wnt/β-catenin pathway. *J. Cancer*. **10**, 5567 (2019).

40. X. Wu, P. Tao, Q. Zhou, J. Li, Z. Yu, X. Wang, J. Li, C. Li, M. Yan, Z. Zhu, B. Liu, L. Su, IL-6 secreted by cancer-associated fibroblasts promotes epithelial-mesenchymal transition and metastasis of gastric cancer via JAK2/STAT3 signaling pathway. *Oncotarget*. **8**, 20741–20750 (2017).

41. L. Wang, L. Cao, H. Wang, B. Liu, Q. Zhang, Z. Meng, X. Wu, Q. Zhou, K. Xu, Cancer-associated fibroblasts enhance metastatic potential of lung cancer cells through IL-6/STAT3 signaling pathway. *Oncotarget*. **8**, 76116–76128 (2017).

42. Y. Yu, C. H. Xiao, L. D. Tan, Q. S. Wang, X. Q. Li, Y. M. Feng, Cancer-associated fibroblasts induce epithelial-mesenchymal transition of breast cancer cells through paracrine TGF-β signalling. *Br. J. Cancer*. **110**, 724–732 (2014).

43. M. Labelle, S. Begum, R. O. Hynes, Direct signaling between platelets and cancer cells induces an epithelial-mesenchymal-like transition and promotes metastasis. *Cancer Cell*. **20**, 576–590 (2011).

44. V. Sigurdsson, B. Hilmarsdottir, H. Sigmundsdottir, A. J. R. Fridriksdottir, M. Ringnér, R. Villadsen, A. Borg, B. A. Agnarsson, O. W. Petersen, M. K. Magnusson, T. Gudjonsson, Endothelial induced EMT in breast epithelial cells with stem cell properties. *PLoS ONE*. **6**, e23833 (2011).

45. Z. Yang, W. Wang, L. Zhao, X. Wang, R. C. Gimple, L. Xu, Y. Wang, J. N. Rich, S. Zhou, Plasma cells shape the mesenchymal identity of ovarian cancers through transfer of exosome-derived microRNAs. *Sci. Adv.* **7** (2021), doi:10.1126/sciadv.abb0737.

46. A. Yadav, B. Kumar, J. Datta, T. N. Teknos, P. Kumar, IL-6 promotes head and neck tumor metastasis by inducing epithelial-mesenchymal transition via the JAK-STAT3-SNAIL signaling pathway. *Mol. Cancer Res.* **9**, 1658–1667 (2011).

47. E. A. Ebbing, A. P. van der Zalm, A. Steins, A. Creemers, S. Hermsen, R. Rentenaar, M. Klein, C. Waasdorp, G. K. J. Hooijer, S. L. Meijer, K. K. Krishnadath, C. J. A. Punt, M. I. van Berge Henegouwen, S. S. Gisbertz, O. M. van Delden, M. C. C. M. Hulshof, J. P. Medema, H. W. M. van Laarhoven, M. F. Bijlsma, Stromal-derived interleukin 6 drives

epithelial-to-mesenchymal transition and therapy resistance in esophageal adenocarcinoma. *Proc Natl Acad Sci USA*. **116**, 2237–2242 (2019).

48. S. Deng, C. Wang, Y. Wang, Y. Xu, X. Li, N. A. Johnson, A. Mukherji, U.-G. Lo, L. Xu, J. Gonzalez, L. A. Metang, J. Ye, C. R. Tirado, K. Rodarte, Y. Zhou, Z. Xie, C. Arana, V. Annamalai, X. Liu, D. J. Vander Griend, P. Mu, Ectopic JAK-STAT activation enables the transition to a stem-like and multilineage state conferring AR-targeted therapy resistance. *Nat. Cancer*. **3**, 1071–1087 (2022).

49. A. H. Sharpe, K. E. Pauken, The diverse functions of the PD1 inhibitory pathway. *Nat. Rev. Immunol.* **18**, 153–167 (2018).

50. Y. Liu, S. Liu, C. Wu, W. Huang, B. Xu, S. Lian, L. Wang, S. Yue, N. Chen, Z. Zhu, PD-1-Mediated PI3K/Akt/mTOR, Caspase 9/Caspase 3 and ERK Pathways Are Involved in Regulating the Apoptosis and Proliferation of CD4+ and CD8+ T Cells During BVDV Infection in vitro. *Front. Immunol.* **11**, 467 (2020).

51. J. Lu, J. Wu, L. Mao, H. Xu, S. Wang, Revisiting PD-1/PD-L pathway in T and B cell response: Beyond immunosuppression. *Cytokine Growth Factor Rev.* **67**, 58–65 (2022).

52. A. Awadasseid, Y. Zhou, K. Zhang, K. Tian, Y. Wu, W. Zhang, Current studies and future promises of PD-1 signal inhibitors in cervical cancer therapy. *Biomed. Pharmacother.* **157**, 114057 (2023).

53. T. Chu, Z. Wang, D. Pe'er, C. G. Danko, Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer*. **3**, 505–517 (2022).

54. Y. Tsujimoto, Role of Bcl-2 family proteins in apoptosis: apoptosomes or mitochondria? *Genes Cells*. **3**, 697–707 (1998).

55. Y. Wang, I. Misumi, A.-D. Gu, T. A. Curtis, L. Su, J. K. Whitmire, Y. Y. Wan, GATA-3 controls the maintenance and proliferation of T cells downstream of TCR and cytokine signaling. *Nat. Immunol.* **14**, 714–722 (2013).

56. L. M. F. Merlo, W. Peng, L. Mandik-Nayak, Impact of IDO1 and IDO2 on the B cell immune response. *Front. Immunol.* **13**, 886225 (2022).

57. A. Bassez, H. Vos, L. Van Dyck, G. Floris, I. Arijs, C. Desmedt, B. Boeckx, M. Vanden Bempt, I. Nevelsteen, K. Lambein, K. Punie, P. Neven, A. D. Garg, H. Wildiers, J. Qian, A. Smeets, D. Lambrechts, A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nat. Med.* **27**, 820–832 (2021).

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

59. S. Hibino, T. Kawazoe, H. Kasahara, S. Itoh, T. Ishimoto, M. Sakata-Yanagimoto, K. Taniguchi, Inflammation-Induced Tumorigenesis and Metastasis. *Int. J. Mol. Sci.* **22** (2021), doi:10.3390/ijms22115421.

60. T. Lu, S. Park, Y. Han, Y. Wang, S. M. Hubert, P. A. Futreal, I. Wistuba, J. V. Heymach, A. Reuben, J. Zhang, T. Wang, Netie: inferring the evolution of neoantigen-T cell interactions in tumors. *Nat. Methods*. **19**, 1480–1489 (2022).

61. T. Lu, S. Wang, L. Xu, Q. Zhou, N. Singla, J. Gao, S. Manna, L. Pop, Z. Xie, M. Chen, J. J. Luke, J. Brugarolas, R. Hannan, T. Wang, Tumor neoantigenicity assessment with CSiN score incorporates clonality and immunogenicity to predict immunotherapy outcomes. *Sci. Immunol.* **5** (2020), doi:10.1126/sciimmunol.aaz3199.

62. M. Sade-Feldman, K. Yizhak, S. L. Bjorgaard, J. P. Ray, C. G. de Boer, R. W. Jenkins, D. J. Lieb, J. H. Chen, D. T. Frederick, M. Barzily-Rokni, S. S. Freeman, A. Reuben, P. J. Hoover, A.-C. Villani, E. Ivanova, A. Portell, P. H. Lizotte, A. R. Aref, J.-P. Eliane, M. R. Hammond, N. Hacohen, Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell*. **175**, 998-1013.e20 (2018).

63. Y. Zhang, H. Chen, H. Mo, X. Hu, R. Gao, Y. Zhao, B. Liu, L. Niu, X. Sun, X. Yu, Y. Wang, Q. Chang, T. Gong, X. Guan, T. Hu, T. Qian, B. Xu, F. Ma, Z. Zhang, Z. Liu, Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. *Cancer Cell*. **39**, 1578-1593.e8 (2021).

64. D. Sun, J. Wang, Y. Han, X. Dong, J. Ge, R. Zheng, X. Shi, B. Wang, Z. Li, P. Ren, L. Sun, Y. Yan, P. Zhang, F. Zhang, T. Li, C. Wang, TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.* **49**, D1420–D1430 (2021).

65. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

66. D. Xiong, Z. Zhang, T. Wang, X. Wang, A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences. *Comput. Struct. Biotechnol. J.* **19**, 3255–3268 (2021).

67. Y. Kim, T. Wang, D. Xiong, X. Wang, S. Park, Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences. *BMC Bioinformatics*. **23**, 469 (2022).

68. S. Park, X. Wang, J. Lim, G. Xiao, T. Lu, T. Wang, Bayesian multiple instance regression for modeling immunogenic neoantigens. *Stat. Methods Med. Res.* **29**, 3032–3047 (2020).

69. G. Fu, X. Nan, H. Liu, R. Y. Patel, P. R. Daga, Y. Chen, D. E. Wilkins, R. J. Doerksen, Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinformatics*. **13 Suppl 15**, S3 (2012).

70. C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods*. **18**, 100–106 (2021).

71. Y. Wang, B. Song, S. Wang, M. Chen, Y. Xie, G. Xiao, L. Wang, T. Wang, Sprod for de-noising spatially resolved transcriptomics data based on position and image information. *Nat. Methods*. **19**, 950–958 (2022).

72. R. Rong, H. Sheng, K. W. Jin, F. Wu, D. Luo, Z. Wen, C. Tang, D. M. Yang, L. Jia, M. Amgad, L. A. D. Cooper, Y. Xie, X. Zhan, S. Wang, G. Xiao, A Deep Learning Approach

for Histology-Based Nucleus Segmentation and Tumor Microenvironment Characterization. *Mod. Pathol.* **36**, 100196 (2023).

73. S. Wang, D. M. Yang, R. Rong, X. Zhan, G. Xiao, Pathology image analysis using segmentation deep learning algorithms. *Am. J. Pathol.* **189**, 1686–1698 (2019).

74. K. Holler, A. Neuschulz, P. Drewe-Boß, J. Mintcheva, B. Spanjaard, R. Arsiè, U. Ohler, M. Landthaler, J. P. Junker, Spatio-temporal mRNA tracking in the early zebrafish embryo. *Nat. Commun.* **12**, 3358 (2021).

75. M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, J. S. Weissman, Molecular recording of mammalian embryogenesis. *Nature*. **570**, 77–82 (2019).

76. A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, J. Shendure, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

77. T. Lu, S. Park, J. Zhu, Y. Wang, X. Zhan, X. Wang, L. Wang, H. Zhu, T. Wang, Overcoming Expressional Drop-outs in Lineage Reconstruction from Single-Cell RNA-Sequencing Data. *Cell Rep.* **34**, 108589 (2021).

78. S. G. Rodriques, L. M. Chen, S. Liu, E. D. Zhong, J. R. Scherrer, E. S. Boyden, F. Chen, RNA timestamps identify the age of single molecules in RNA sequencing. *Nat. Biotechnol.* **39**, 320–325 (2021).

79. R. U. Sheth, H. H. Wang, DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* **19**, 718–732 (2018).

80. B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nuñez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, J. S. Weissman, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*. **167**, 1867-1882.e21 (2016).

81. A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, A. Regev, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*. **167**, 1853-1866.e17 (2016).

82. Z. Zhang, D. Xiong, X. Wang, H. Liu, T. Wang, Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods*. **18**, 92–99 (2021).

83. Z. Zhang, W. Y. Chang, K. Wang, Y. Yang, X. Wang, C. Yao, T. Wu, L. Wang, T. Wang, Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. *Nat. Mach. Intell.* (2022), doi:10.1038/s42256-022-00492-6.

84. S. Gogola, M. Rejzer, H. F. Bahmad, W. Abou-Kheir, Y. Omarzai, R. Poppiti, Epithelial-to-Mesenchymal Transition-Related Markers in Prostate Cancer: From Bench to Bedside. *Cancers (Basel)*. **15** (2023), doi:10.3390/cancers15082309.

85. G. H. Henry, A. Malewska, D. B. Joseph, V. S. Malladi, J. Lee, J. Torrealba, R. J. Mauck, J. C. Gahan, G. V. Raj, C. G. Roehrborn, G. C. Hon, M. P. MacConmara, J. C. Reese, R. C. Hutchinson, C. M. Vezina, D. W. Strand, A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell Rep.* **25**, 3530-3542.e5 (2018).

86. H. Song, H. N. W. Weinstein, P. Allegakoen, M. H. Wadsworth, J. Xie, H. Yang, E. A. Castro, K. L. Lu, B. A. Stohr, F. Y. Feng, P. R. Carroll, B. Wang, M. R. Cooperberg, A. K. Shalek, F. W. Huang, Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nat. Commun.* **13**, 141 (2022).

87. T. Lu, Z. Zhang, J. Zhu, Y. Wang, P. Jiang, X. Xiao, C. Bernatchez, J. V. Heymach, D. L. Gibbons, J. Wang, L. Xu, A. Reuben, T. Wang, Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).

Fig. 1

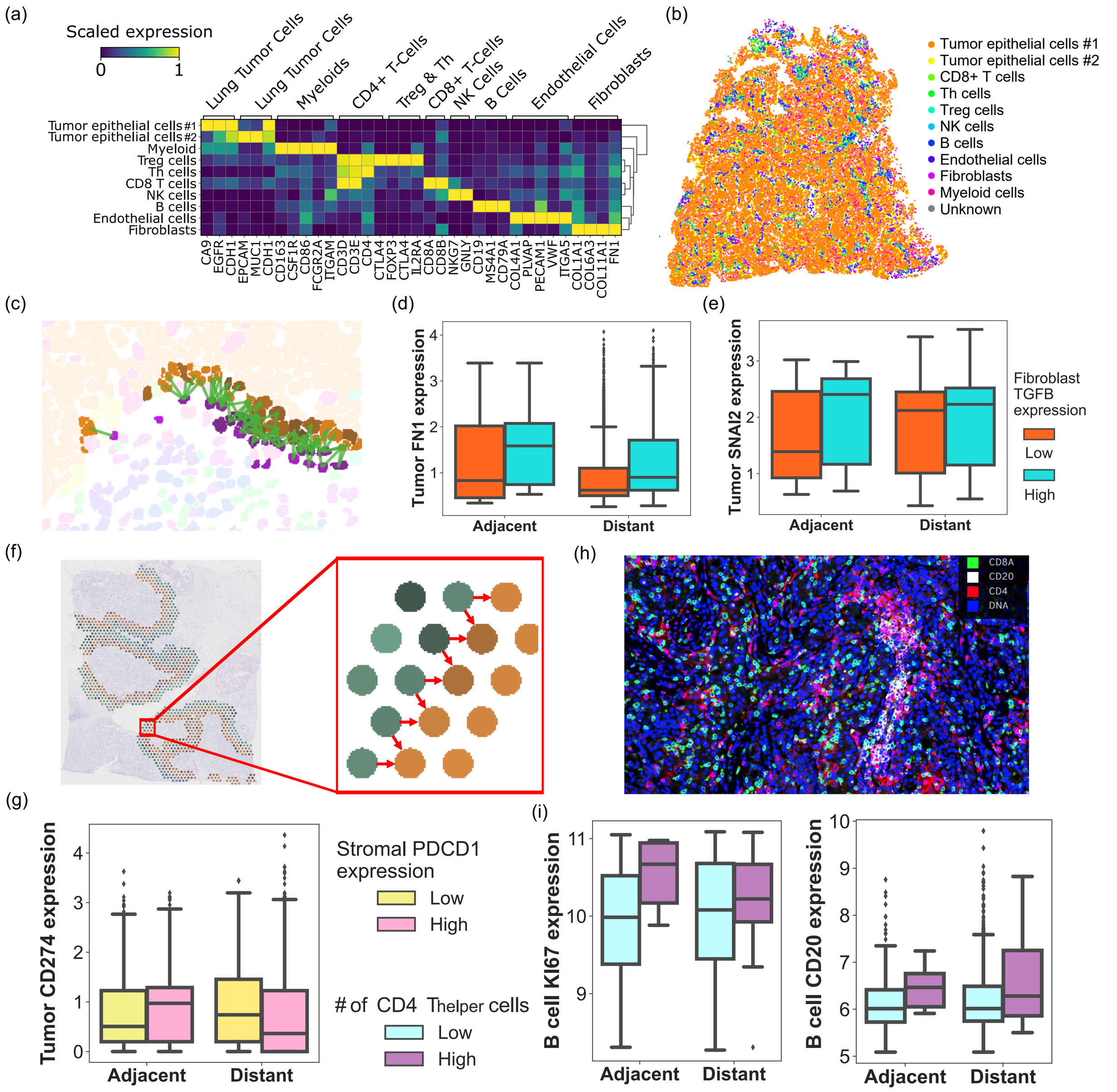

(a)
Scaled expression
0
1
Lung Tumor Cells
Lung Tumor Cells
Myeloids
CD4+ T-Cells
Treg & Th
CD8+ T-Cells
NK Cells
B Cells
Endothelial Cells
Fibroblasts
Tumor epithelial cells #1
Tumor epithelial cells #2
Myeloid
Treg cells
Th cells
CD8 T cells
NK cells
B cells
Endothelial cells
Fibroblasts
CA9 EGFR CDH1 EPCAM MUC1 CDH1 CD163 CSF1R CD86 FCGR2A ITGAM CD3D CD3E CD4 CTLA4 FOXP3 CTLA4 IL2RA CD8A CD8B NKG7 GNLY CD19 MS4A1 CD79A COL4A1 PLVAP PECAM1 VWF ITGA5 COL1A1 COL6A3 COL11A1 FN1
(b)
Tumor epithelial cells #1
Tumor epithelial cells #2
CD8+ T cells
Th cells
Treg cells
NK cells
B cells
Endothelial cells
Fibroblasts
Myeloid cells
Unknown
(c)
(d)
Tumor FN1 expression
Adjacent
Distant
(e)
Tumor SNAI2 expression
Adjacent
Distant
Fibroblast TGFB expression
Low
High
(f)
(h)
CD8A
CD20
CD4
DNA
(g)
Tumor CD274 expression
Adjacent
Distant
Stromal PDCD1 expression
Low
High
# of CD4 Thelper cells
Low
High
(i)
B cell KI67 expression
Adjacent
Distant
B cell CD20 expression
Adjacent
Distant

Fig. 2

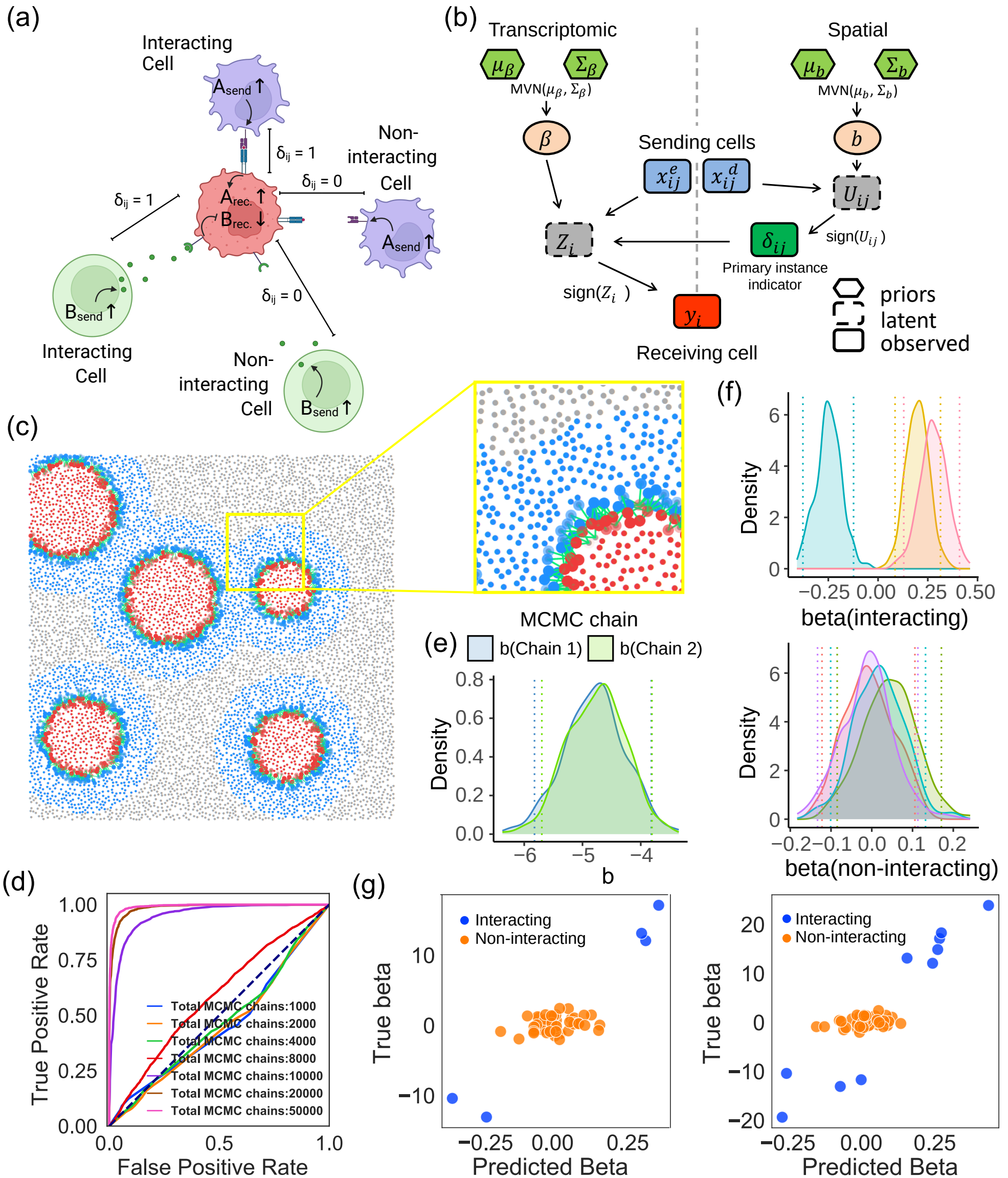

(a)
Interacting Cell
Non-interacting Cell
Interacting Cell
Non-interacting Cell
(b)
Transcriptomic
Spatial
Sending cells
Primary instance indicator
Receiving cell
priors
latent
observed
(c)
(d)
True Positive Rate
False Positive Rate
Total MCMC chains:1000
Total MCMC chains:2000
Total MCMC chains:4000
Total MCMC chains:8000
Total MCMC chains:10000
Total MCMC chains:20000
Total MCMC chains:50000
(e)
MCMC chain
b(Chain 1)
b(Chain 2)
Density
b
(f)
Density
beta(interacting)
beta(non-interacting)
(g)
Interacting
Non-interacting
True beta
Predicted Beta

Fig. 3

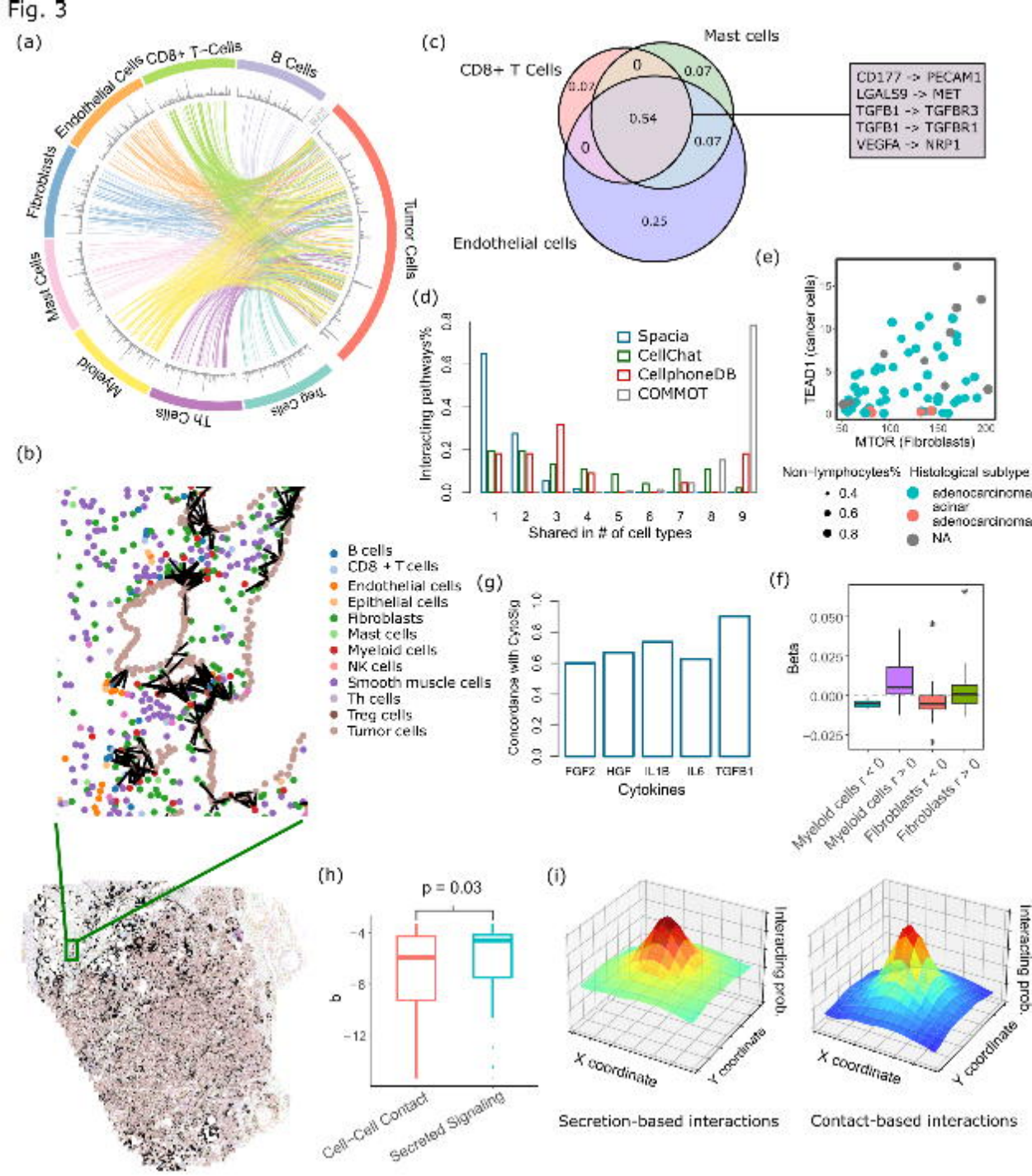

(a)
CD8+ T-Cells
B Cells
Endothelial Cells
Fibroblasts
Mast Cells
Myeloid
Th Cells
Treg Cells
Tumor Cells
(b)
B cells
CD8 + T cells
Endothelial cells
Epithelial cells
Fibroblasts
Mast cells
Myeloid cells
NK cells
Smooth muscle cells
Th cells
Treg cells
Tumor cells
(c)
Mast cells
CD8+ T Cells
Endothelial cells
0
0.07
0.07
0.54
0
0.07
0.25
CD177 -> PECAM1
LGALS9 -> MET
TGFB1 -> TGFBR3
TGFB1 -> TGFBR1
VEGFA -> NRP1
(d)
Spacia
CellChat
CellphoneDB
COMMOT
Interacting pathways%
Shared in # of cell types
(e)
TEAD1 (cancer cells)
MTOR (Fibroblasts)
Non-lymphocytes%
0.4
0.6
0.8
Histological subtype
adenocarcinoma
acinar adenocarcinoma
NA
(f)
Beta
Myeloid cells r < 0
Myeloid cells r > 0
Fibroblasts r < 0
Fibroblasts r > 0
(g)
Concordance with CytoSig
FGF2
HGF
IL1B
IL6
TGFB1
Cytokines
(h)
p = 0.03
b
Cell-Cell Contact
Secreted Signaling
(i)
Interacting prob.
X coordinate
Y coordinate
Secretion-based interactions
Contact-based interactions

Fig. 4

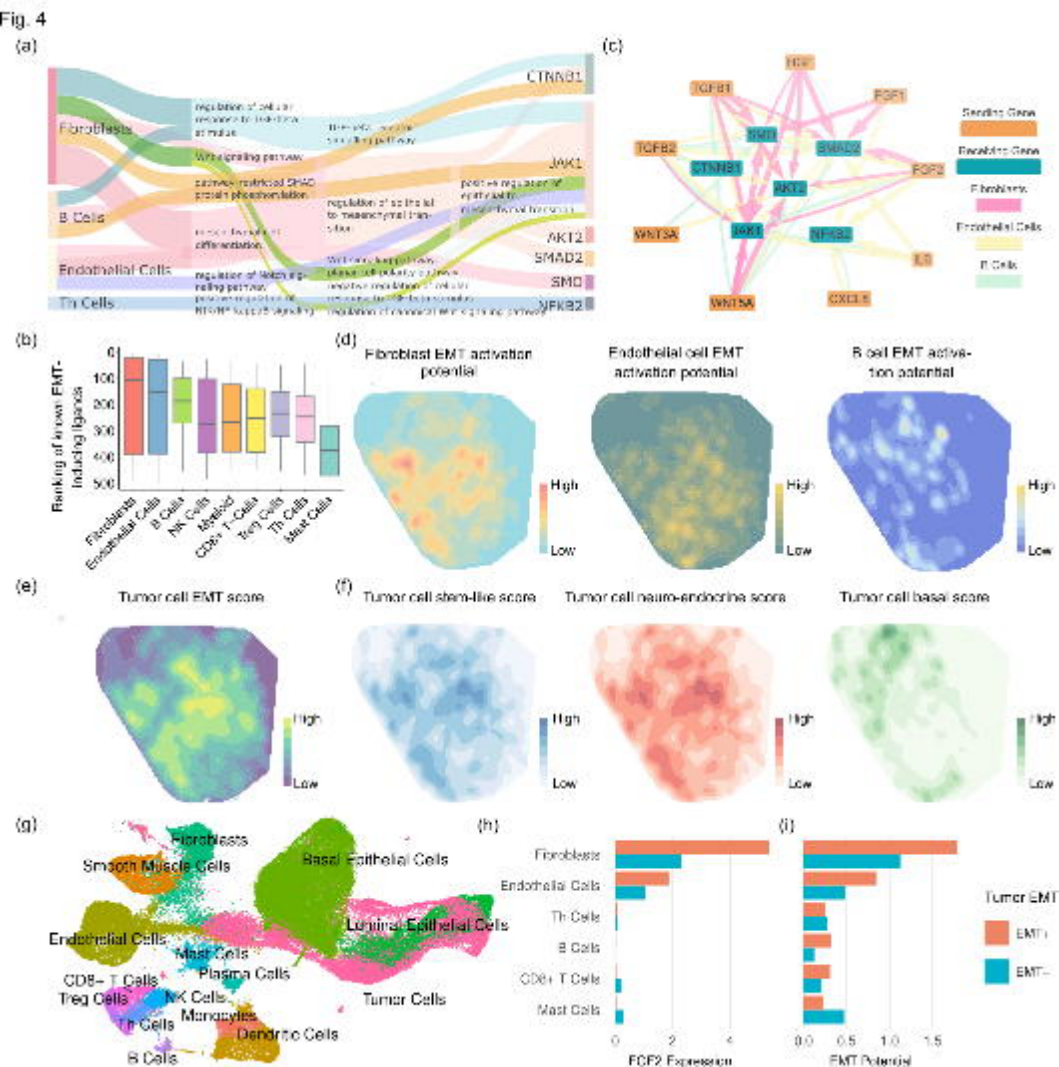

(a)
Fibroblasts
B Cells
Endothelial Cells
Th Cells
CTNNB1
JAK1
AKT2
SMAD2
SMO
NFKB2
(b)
Ranking of known EMT-inducing ligands
Fibroblasts
Endothelial Cells
B Cells
NK Cells
Myeloid
CD8+ T Cells
Treg Cells
Th Cells
Mast Cells
(c)
HGF
TGFB1
FGF1
SMO
SMAD2
TGFB2
CTNNB1
FGF2
AKT2
WNT5A
JAK1
NFKB2
IL6
WNT5A
CXCL8
Sending Gene
Receiving Gene
Fibroblasts
Endothelial Cells
B Cells
(d)
Fibroblast EMT activation potential
Endothelial cell EMT activation potential
B cell EMT activation potential
High
Low
(e)
Tumor cell EMT score
(f)
Tumor cell stem-like score
Tumor cell neuro-endocrine score
Tumor cell basal score
(g)
Fibroblasts
Smooth Muscle Cells
Basal Epithelial Cells
Luminal Epithelial Cells
Endothelial Cells
Mast Cells
Plasma Cells
CD8+ T Cells
Treg Cells
NK Cells
Tumor Cells
Th Cells
Monocytes
Dendritic Cells
B Cells
(h)
Fibroblasts
Endothelial Cells
Th Cells
B Cells
CD8+ T Cells
Mast Cells
FGF2 Expression
(i)
EMT Potential
Tumor EMT
EMT+
EMT−

Fig. 5

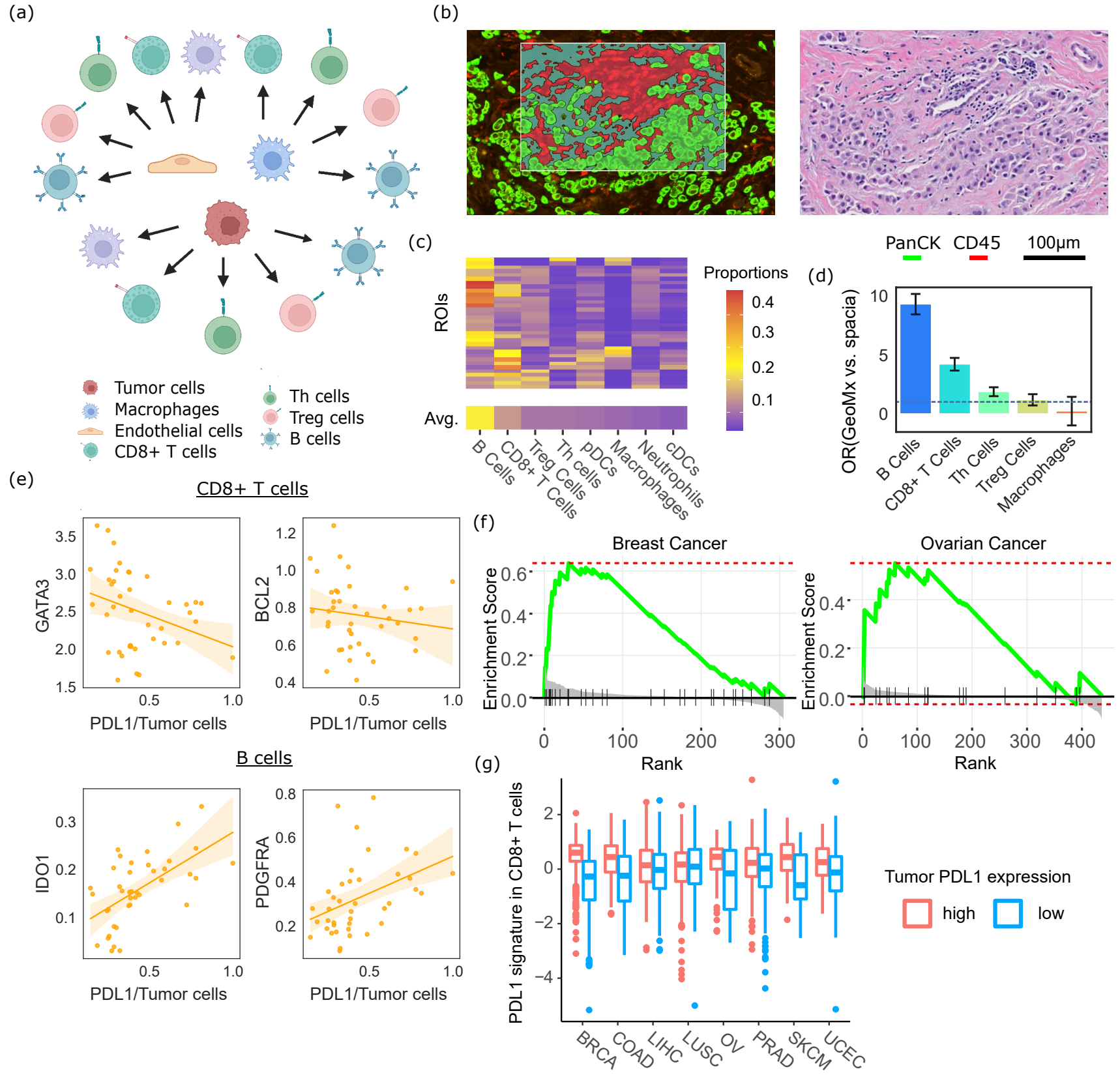
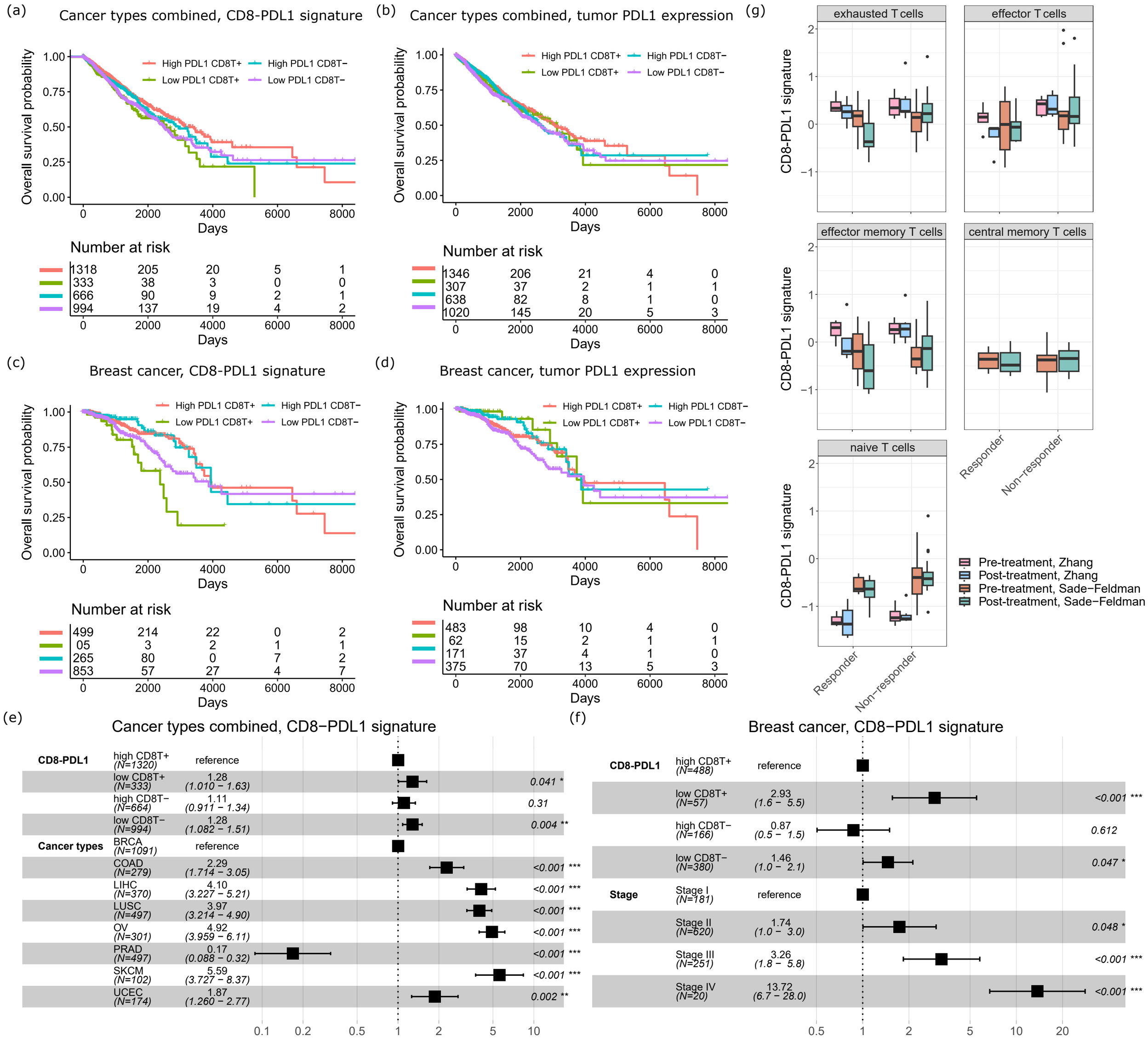
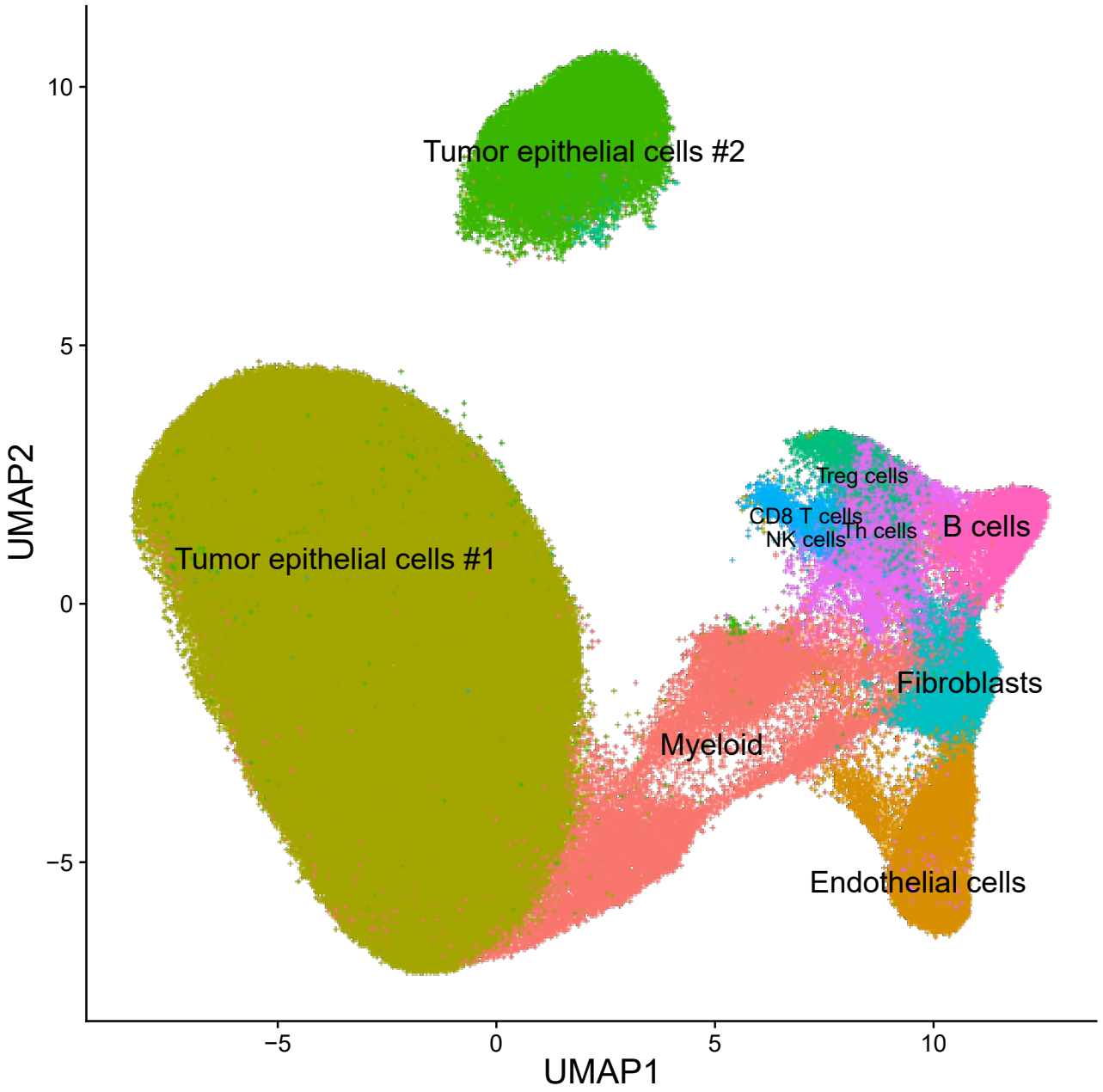

(a)
Tumor cells
Macrophages
Endothelial cells
CD8+ T cells
Th cells
Treg cells
B cells
(b)
PanCK
CD45
100μm
(c)
ROIs
Avg.
Proportions
0.4
0.3
0.2
0.1
B Cells
CD8+ T Cells
Treg Cells
Th cells
pDCs
Macrophages
Neutrophils
cDCs
(d)
OR(GeoMx vs. spacia)
10
5
0
B Cells
CD8+ T Cells
Th Cells
Treg Cells
Macrophages
(e)
CD8+ T cells
GATA3
BCL2
PDL1/Tumor cells
B cells
IDO1
PDGFRA
(f)
Breast Cancer
Ovarian Cancer
Enrichment Score
Rank
(g)
PDL1 signature in CD8+ T cells
BRCA
COAD
LIHC
LUSC
OV
PRAD
SKCM
UCEC
Tumor PDL1 expression
high
low

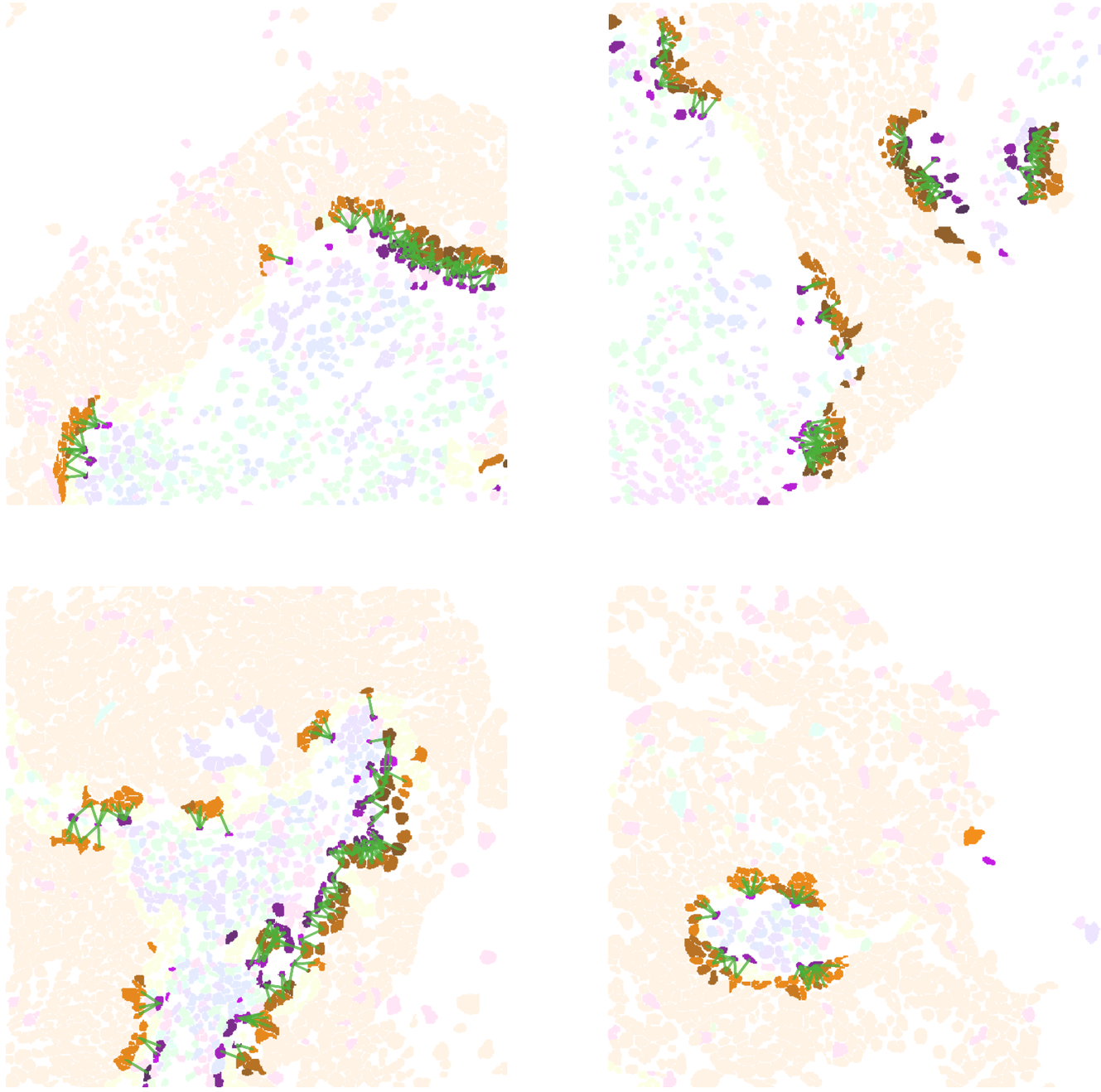
Fig. 6


(a)
Cancer types combined, CD8-PDL1 signature
Overall survival probability
High PDL1 CD8T+
High PDL1 CD8T−
Low PDL1 CD8T+
Low PDL1 CD8T−
Days
Number at risk
1318 205 20 5 1
333 38 3 0 0
666 90 9 2 1
994 137 19 4 2
(b)
Cancer types combined, tumor PDL1 expression
Number at risk
1346 206 21 4 0
307 37 2 1 1
638 82 8 1 0
1020 145 20 5 3
(c)
Breast cancer, CD8-PDL1 signature
Number at risk
499 214 22 0 2
05 3 2 1 1
265 80 0 7 2
853 57 27 4 7
(d)
Breast cancer, tumor PDL1 expression
Number at risk
483 98 10 4 0
62 15 2 1 1
171 37 4 1 0
375 70 13 5 3
(g)
exhausted T cells
effector T cells
effector memory T cells
central memory T cells
naive T cells
CD8-PDL1 signature
Responder
Non−responder
Pre-treatment, Zhang
Post-treatment, Zhang
Pre-treatment, Sade−Feldman
Post-treatment, Sade−Feldman
(e)
Cancer types combined, CD8−PDL1 signature
CD8-PDL1
high CD8T+ (N=1320) reference
low CD8T+ (N=333) 1.28 (1.010 − 1.63) 0.041 *
high CD8T− (N=664) 1.11 (0.911 − 1.34) 0.31
low CD8T− (N=994) 1.28 (1.082 − 1.51) 0.004 **
Cancer types
BRCA (N=1091) reference
COAD (N=279) 2.29 (1.714 − 3.05) <0.001 ***
LIHC (N=370) 4.10 (3.227 − 5.21) <0.001 ***
LUSC (N=497) 3.97 (3.214 − 4.90) <0.001 ***
OV (N=301) 4.92 (3.959 − 6.11) <0.001 ***
PRAD (N=497) 0.17 (0.088 − 0.32) <0.001 ***
SKCM (N=102) 5.59 (3.727 − 8.37) <0.001 ***
UCEC (N=174) 1.87 (1.260 − 2.77) 0.002 **
(f)
Breast cancer, CD8−PDL1 signature
CD8-PDL1
high CD8T+ (N=488) reference
low CD8T+ (N=57) 2.93 (1.6 − 5.5) <0.001 ***
high CD8T− (N=166) 0.87 (0.5 − 1.5) 0.612
low CD8T− (N=380) 1.46 (1.0 − 2.1) 0.047 *
Stage
Stage I (N=181) reference
Stage II (N=620) 1.74 (1.0 − 3.0) 0.048 *
Stage III (N=251) 3.26 (1.8 − 5.8) <0.001 ***
Stage IV (N=20) 13.72 (6.7 − 28.0) <0.001 ***

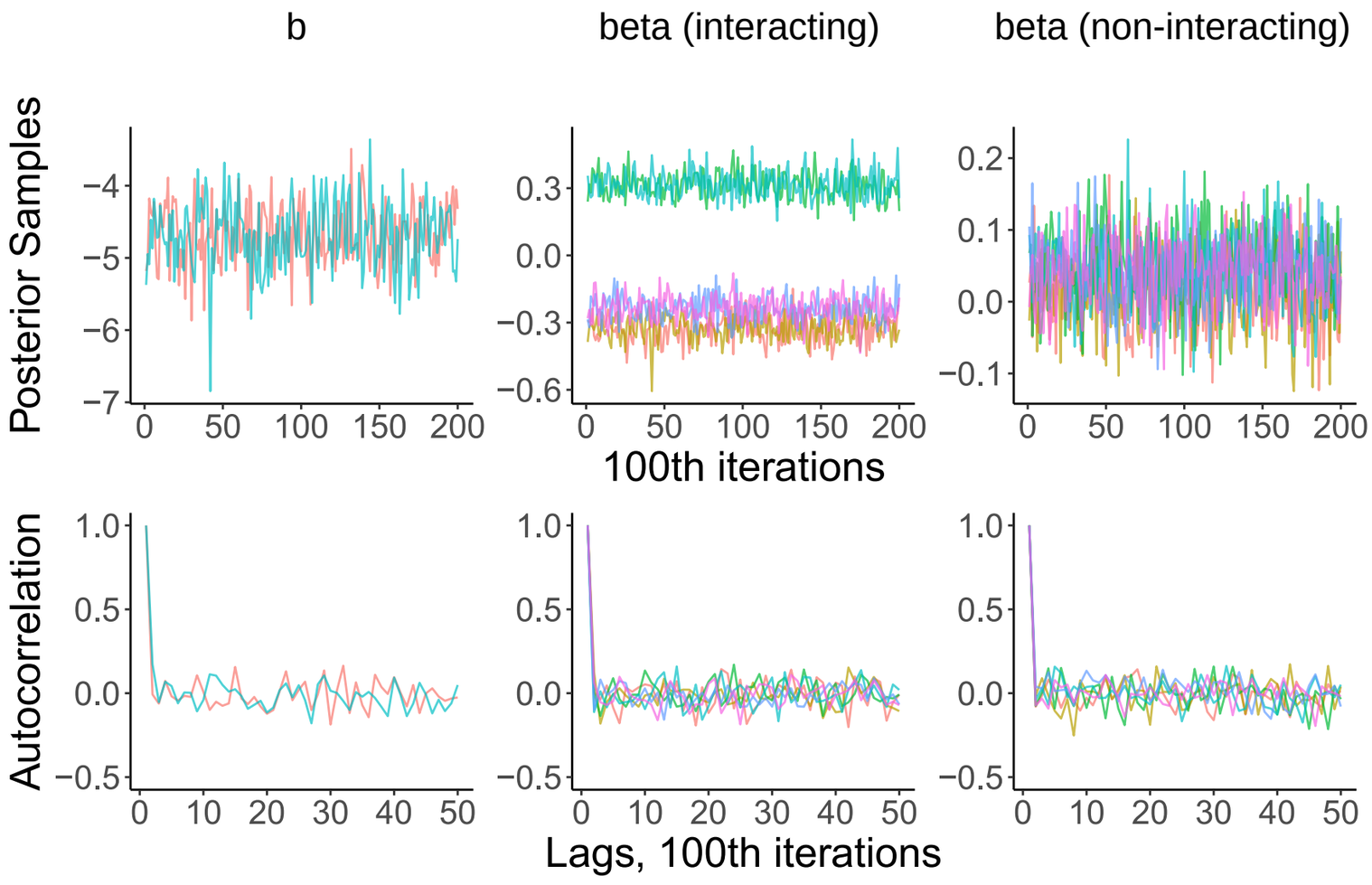
Sup. Fig. 1

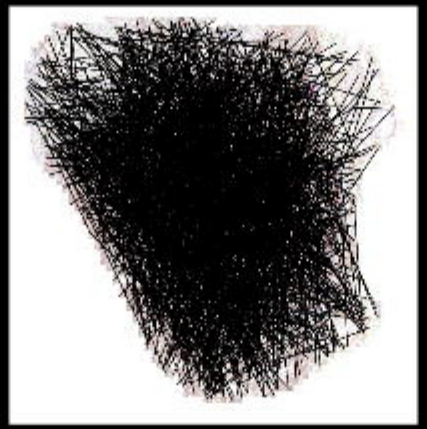

Tumor epithelial cells #2
Tumor epithelial cells #1
Treg cells
CD8 T cells
NK cells
Th cells
B cells
Fibroblasts
Myeloid
Endothelial cells
UMAP2
UMAP1
10
5
0
−5
−5
0
5
10

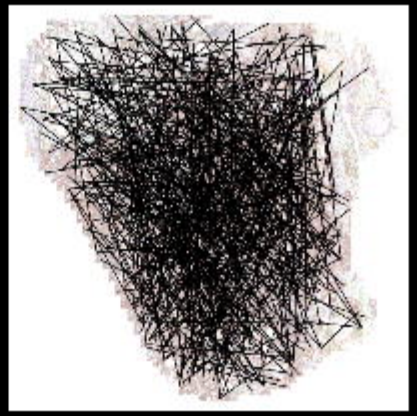
Sup. Fig. 2



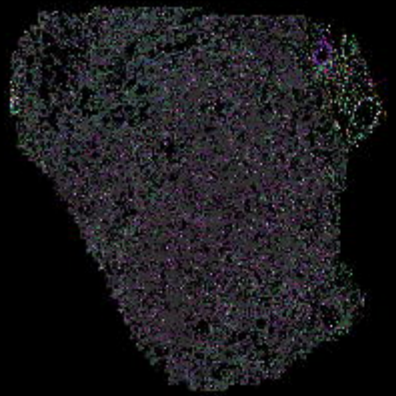
Sup. Fig. 3


b
beta (interacting)
beta (non-interacting)
Posterior Samples
−4
−5
−6
−7
0.3
0.0
−0.3
−0.6
0.2
0.1
0.0
−0.1
0
50
100
150
200
100th iterations
Autocorrelation
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
Lags, 100th iterations

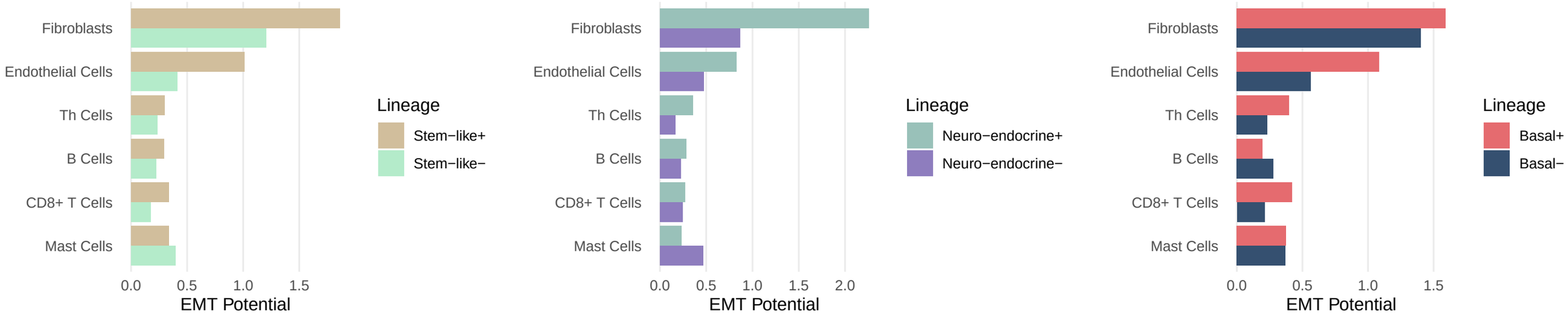




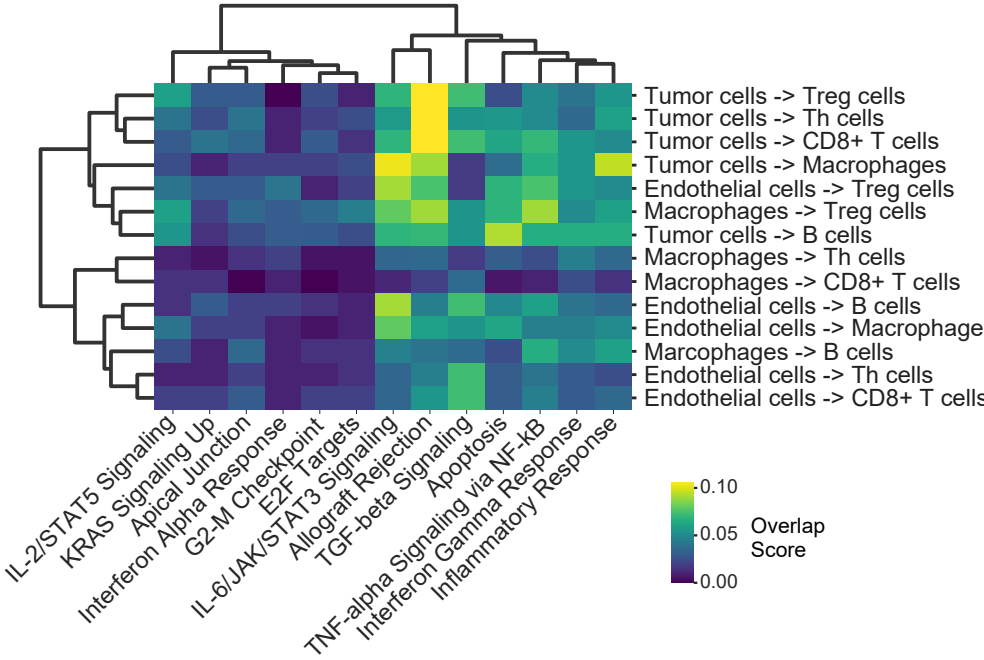
- B cells
- CD8+ T cells
- Endothelial cells
- Normal epithelial cells
- Fibroblasts
- Mast cells
- Myeloid cells
- NK cells
- Smooth muscle cells
- Th cells
- Treg cells
- Tumor epithelial cells



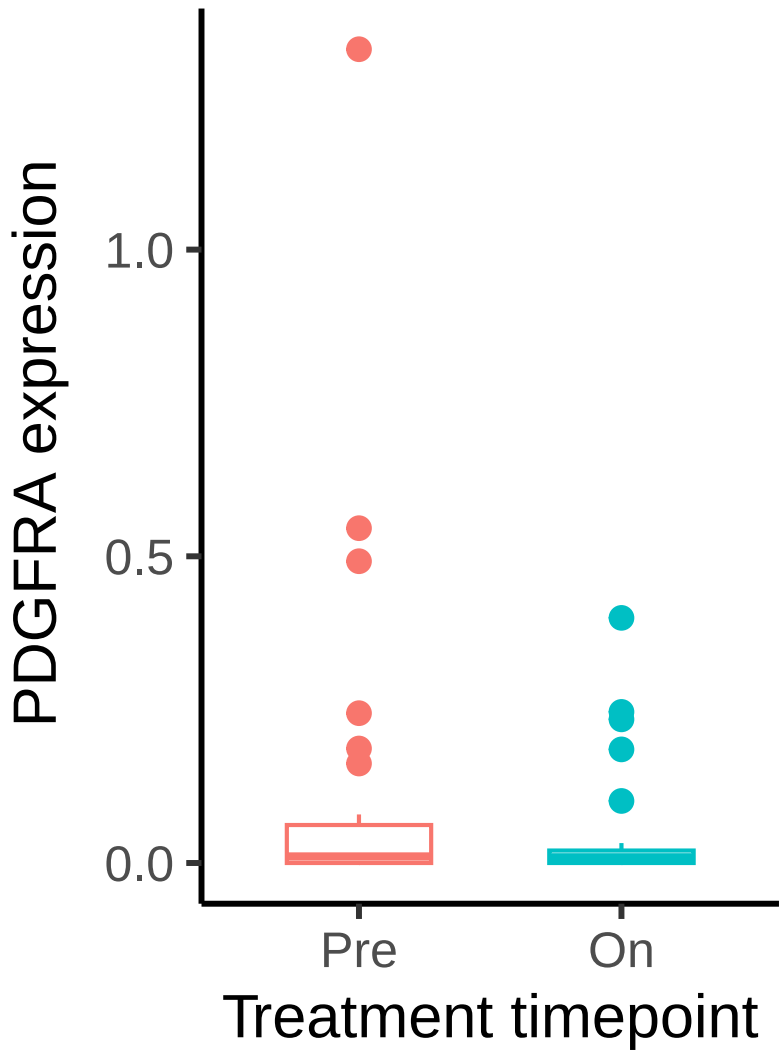
- B_cells
- CD8_T-cells
- Endothelial_cells
- Epithelial_Cells
- Fibroblasts
- Mast_cells
- Myeloid_cells
- NK_cells
- Smooth_Muscle
- Th_cells
- Treg_cells
- Tumor_cells
- Unknown_cells


Fibroblasts
Endothelial Cells
Th Cells
B Cells
CD8+ T Cells
Mast Cells
0.0
0.5
1.0
1.5
EMT Potential
Lineage
Stem−like+
Stem−like−
Fibroblasts
Endothelial Cells
Th Cells
B Cells
CD8+ T Cells
Mast Cells
0.0
0.5
1.0
1.5
2.0
EMT Potential
Lineage
Neuro−endocrine+
Neuro−endocrine−
Fibroblasts
Endothelial Cells
Th Cells
B Cells
CD8+ T Cells
Mast Cells
0.0
0.5
1.0
1.5
EMT Potential
Lineage
Basal+
Basal−

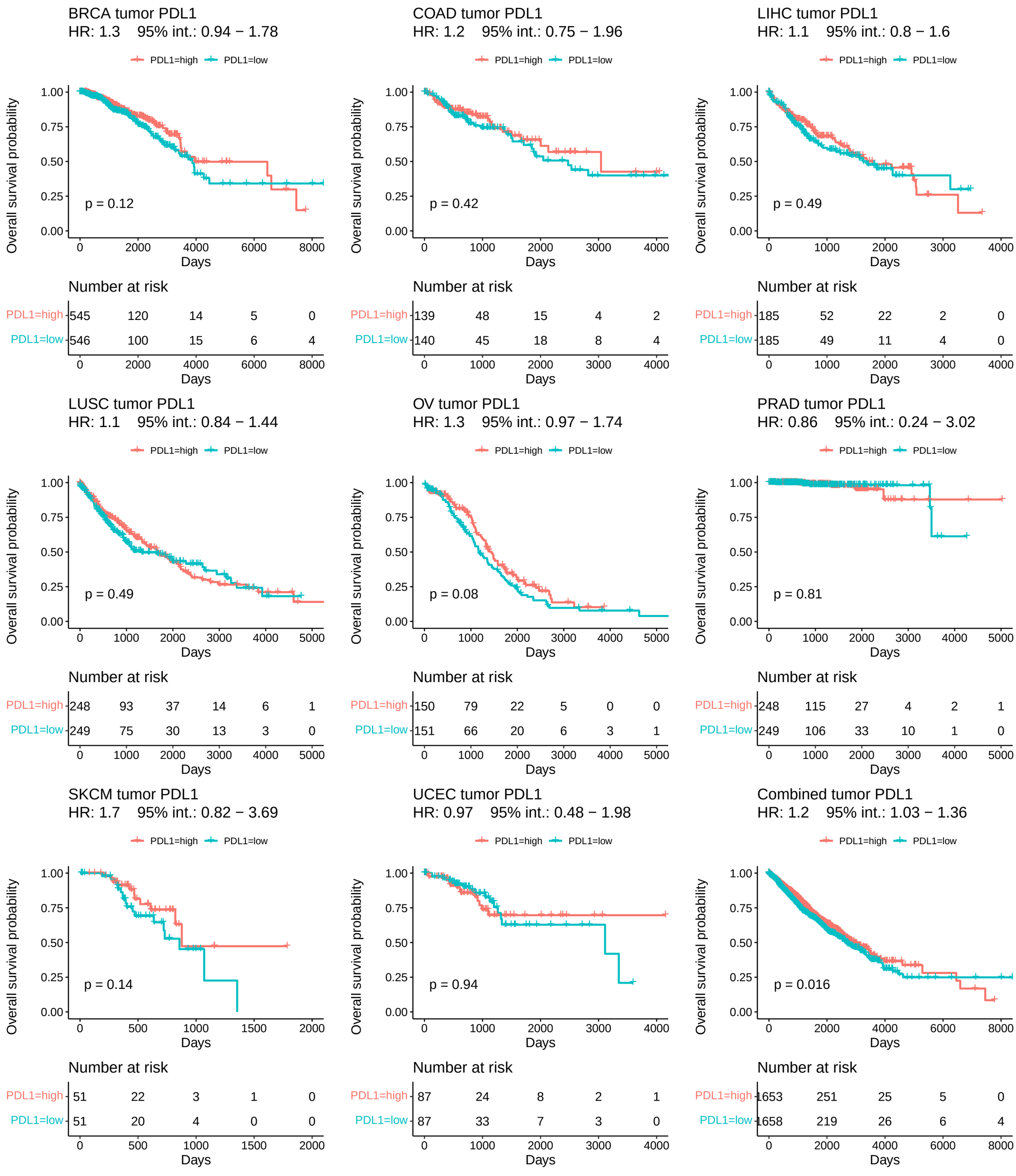
Sup. Fig. 6


Tumor cells -> Treg cells
Tumor cells -> Th cells
Tumor cells -> CD8+ T cells
Tumor cells -> Macrophages
Endothelial cells -> Treg cells
Macrophages -> Treg cells
Tumor cells -> B cells
Macrophages -> Th cells
Macrophages -> CD8+ T cells
Endothelial cells -> B cells
Endothelial cells -> Macrophages
Marcophages -> B cells
Endothelial cells -> Th cells
Endothelial cells -> CD8+ T cells
IL-2/STAT5 Signaling
KRAS Signaling Up
Apical Junction
Interferon Alpha Response
G2-M Checkpoint
E2F Targets
IL-6/JAK/STAT3 Signaling
Allograft Rejection
TGF-beta Signaling
Apoptosis
TNF-alpha Signaling via NF-kB
Interferon Gamma Response
Inflammatory Response
0.10
0.05
0.00
Overlap Score

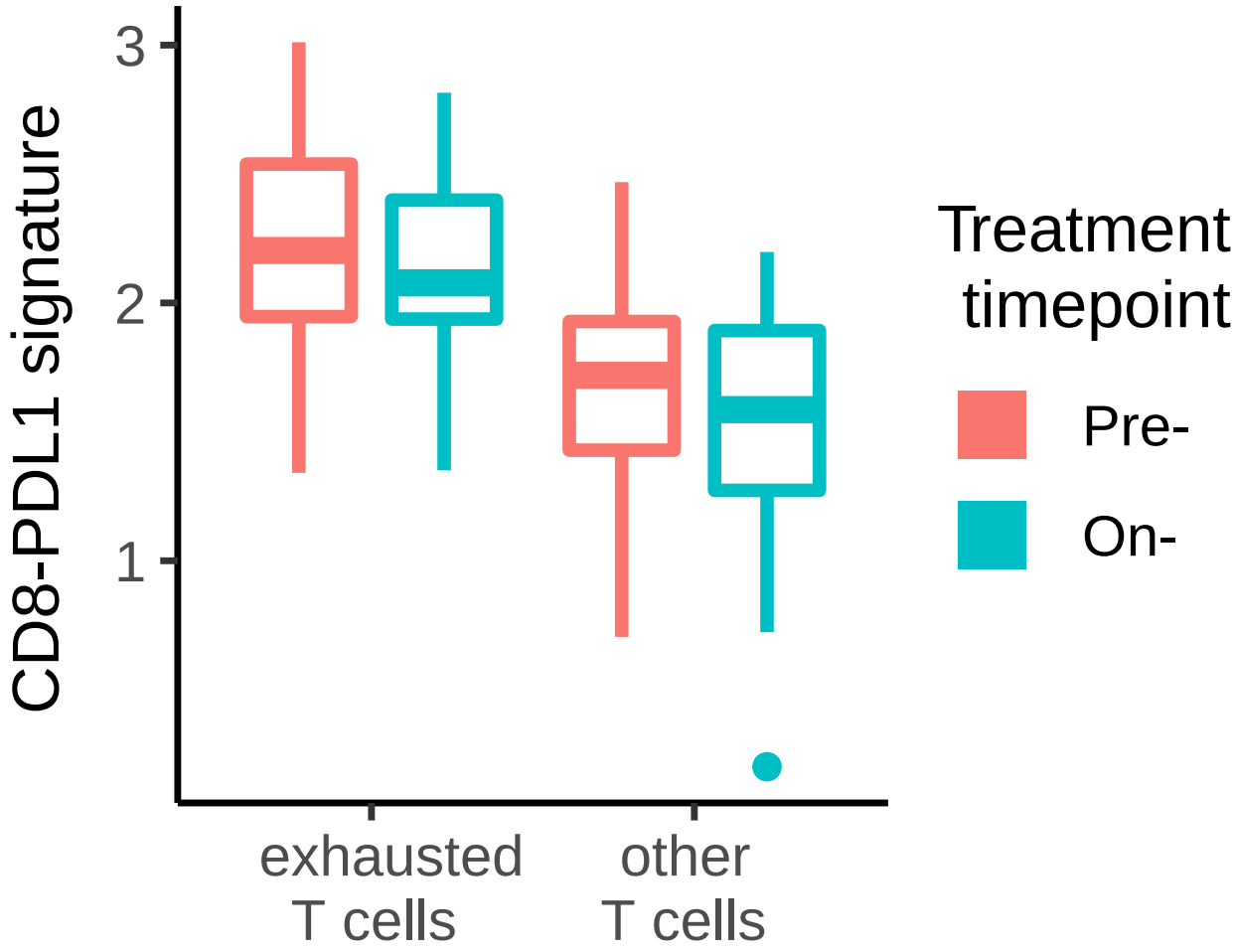
Sup. Fig. 7



Sup. Fig. 8


BRCA tumor PDL1
HR: 1.3 95% int.: 0.94 − 1.78
COAD tumor PDL1
HR: 1.2 95% int.: 0.75 − 1.96
LIHC tumor PDL1
HR: 1.1 95% int.: 0.8 − 1.6
LUSC tumor PDL1
HR: 1.1 95% int.: 0.84 − 1.44
OV tumor PDL1
HR: 1.3 95% int.: 0.97 − 1.74
PRAD tumor PDL1
HR: 0.86 95% int.: 0.24 − 3.02
SKCM tumor PDL1
HR: 1.7 95% int.: 0.82 − 3.69
UCEC tumor PDL1
HR: 0.97 95% int.: 0.48 − 1.98
Combined tumor PDL1
HR: 1.2 95% int.: 1.03 − 1.36
PDL1=high
PDL1=low
Overall survival probability
Days
p = 0.12
p = 0.42
p = 0.49
p = 0.49
p = 0.08
p = 0.81
p = 0.14
p = 0.94
p = 0.016
Number at risk
PDL1=high 545 120 14 5 0
PDL1=low 546 100 15 6 4
PDL1=high 139 48 15 4 2
PDL1=low 140 45 18 8 4
PDL1=high 185 52 22 2 0
PDL1=low 185 49 11 4 0
PDL1=high 248 93 37 14 6 1
PDL1=low 249 75 30 13 3 0
PDL1=high 150 79 22 5 0 0
PDL1=low 151 66 20 6 3 1
PDL1=high 248 115 27 4 2 1
PDL1=low 249 106 33 10 1 0
PDL1=high 51 22 3 1 0
PDL1=low 51 20 4 0 0
PDL1=high 87 24 8 2 1
PDL1=low 87 33 7 3 0
PDL1=high 1653 251 25 5 0
PDL1=low 1658 219 26 6 4

Sup. Fig. 9


CD8-PDL1 signature
3
2
1
exhausted
T cells
other
T cells
Treatment
timepoint
Pre-
On-