SPRUCE: Single-cell Pairwise Relationships Untangled by Composite Embedding model

Sishir Subedi and Yongjin P. Park

11:38:35 AM, Sep 16, 2022

Summary

In multi-cellular organisms, cell identity and functions are primed and refined through interactions with other surrounding cells. Here, we propose a scalable machine learning method, termed SPURCE, which is designed to systematically ascertain common cell-cell communication patterns embedded in single-cell RNA-seq data. We applied our approach to investigate tumour microenvironments consolidating multiple breast cancer data sets and found seven frequently-observed interaction signatures and underlying gene-gene interaction networks. Our results implicate that a part of tumour heterogeneity, especially within the same subtype, is better understood by differential interaction patterns rather than the static expression of known marker genes.

Keywords

Single-cell genomics, network, cellular communication, surface protein interaction, deep learning, variational autoencoder, probabilistic topic modelling

Introduction

The advancement in single-cell RNA-sequencing (scRNA-seq) has emerged as a new frontier in genomics. Quantification of multimodal omics at a single-cell resolution has made it possible to gain insights into different aspects of cancer biology [Teichmann and Efremova, 2020]. One of the fundamental questions in cancer research is how cancer cells interact with each other in a confined heterogeneous environment such as a tumour microenvironment (TME). Studies have shown that cell-cell communication (CCC) among cell populations in the TME is crucial in cancer growth and metastatic processes [Tan and Naylor, 2022]. Understanding the intricacies of communication among tumour and their interacting partner cells could aid in identifying a potential therapeutic avenue in cancer.

A significant technical challenge that stands in understanding the dynamics of cell-cell interactions in TME is devising a systematic approach to isolate and capture interaction signals from each interacting cell pair. A conventional approach to studying CCC involves clustering features in low-dimensional space and inferring interactions between clusters of know cell types [Almet et al., 2021, Jin et al., 2021, Efremova et al., 2020]. While these methods have uncovered numerous signalling mechanisms that govern cellular differentiation and pathogenesis, they assume each cluster, annotated using a limited number of marker genes, represents a cell type; hence all the cells within

a cluster interact in the same manner. These methods do not account for intracluster cellular heterogeneity. Cells within a cell type may exist in multiple subtypes/states and manifest heterogeneous interaction patterns based on the type and state of the interacting partner cell, which is critical in understanding cancer progression [Zhang et al., 2021a, tan2022tumour]. Additionally, interaction among cells in different contexts, such as disease states, are studied separately, which loses context-specific variability information and are repetitive and computationally expensive.

Recent studies have addressed these challenges and developed methods to capture the diversity of cell interactions within the same cluster. `Tensor-cell2cell` [Armingol et al., 2022] uses tensor-based dimensionality reduction techniques to infer context-driven CCC pattern. `scTensor` [Tsuyuzaki et al., 2019] also uses a tensor decomposition algorithm to infer many-to-many cell-pair relationships as a hypergraph. These methods rely on a priori knowledge of cell type and aggregating cells to calculate communication scores based on the mean expression of ligand-receptor (LR) genes. `SoptSC` [Wang et al., 2019] calculates signalling probability between two cells based on pathway-specific LR and target genes and addresses heterogeneity of cells within the same cluster. However, the method requires a user-defined comprehensive list of pathway genes and does not scale to cohort-level studies.

Here, we present a novel computational approach termed SPRUCE, Single-cell Pairwise Relationship Untangled by Composite Embedding, to analyze tens of millions of cell pairs in a scalable way. Adopting known ligand and receptor protein-protein interactions, we asked how and why cell pairs are localized in the proximity of latent topic space and highlighted common patterns repeatedly observed in tumour microenvironments data. SPRUCE is based on an embedded topic model (ETM), a generative deep learning method built on variational autoencoder architecture, and represents single-cell vector data in low-dimension topic space with an interpretable topic-specific gene expression dictionary matrix. It has been successfully implemented in natural language processing to extract meaningful topics representing large-scale documents [Dieng et al., 2020]. A recent study, scETM, showed that ETM-based techniques efficiently capture essential biological signals from sparse and heterogeneous single-cell data [Zhao et al., 2021]. The key contribution of our approach is the unbiased identification of interpretable cell subtype/state across multiple datasets by characterizing LR genes-driven patterns of cell-cell interactions. Existing graph-based single-cell analysis methods often define cell-cell interaction modules as densely-connected components in a graph (an adjacency matrix). Our SPRUCE model considers cell-cell interaction patterns as a stream of edges, or a giant incidence matrix (edge by vertex or other vertex property).

Results

Overview of SPRUCE model training in breast cancer study

We combined existing breast cancer data sets [Wu et al., 2021, Bhat-Nakshatri et al., 2021] and cancer-specific immune cell data [Zheng et al., 2021] and constructed a comprehensive single-cell catalogue for an unbiased breast cancer study, yielding a data matrix consisting of 20,288 genes and 155,913 cells. We first mapped cells from multiple data sets onto a common latent topic space (K=50) and harmonized them based on variational autoencoder-based topic modelling, not posing additional assumptions, such as a selection bias imposed by top marker genes. Based on cosine similarity in the latent topic space between cells, we then constructed cell-cell interaction networks and performed stratified sampling so that topic-topic relationships are similarly represented in the subsequent steps in SPRUCE training. For each of these 25M+ cell pairs, we extracted gene expressions of the known 648 ligand and 672 receptor proteins [Shao et al., 2021] and used them as

feature vectors for SPRUCE model training.

Multinomial probabilistic topic modelling identified 50 cell topics across 11 known cell types

We implemented a Bayesian deep learning approach to estimate embedded topic models across 155,913 cells with 50 latent dimensions. We found each cell topic corresponds to a group of an average of 3118 cells (with a standard deviation of $\pm$ 5,987) (Figure 1A). Among 50 topics, 32 of them contained more than 100 cells. The highest number of cells (24% of the dataset) were assigned to topic 37, in which 96% of cells were previously identified immune cells (T and B cells). 98% of cancer cells from the dataset were assigned to 13 cell topics. The cancer cell proportion in 9 of 13 topics was greater than 95%. The latent cell topics with cell type annotated were visualized with UMAP, showing distinct clusters for each topic where the majority of cells belong to one of the major types of cells in the dataset (Figure 1B). We further confirmed concordance with the previous analysis conducted in original papers using cell type lineage canonical markers (Figure 1D). The estimated cell topic proportions show that the resident cell types have similar topic proportions. However, cancer cells have a different mixture of topic proportions which shows that the model identified many distinct topics of cancer cells (Figure 1E). We also tried a different number of cell topics from 10, 25, and 50 and decided to use the 50-topic model because major cell types, especially cancer cells, showed well-separated distinct clusters.

Seven robust TME-specific interaction signatures were found in 25 million cell-cell pairs.

We constructed LR gene expression data from 155,913 cells to construct a set of 24,790,167 cell pairs and estimated embedded interaction topic models with 25 latent dimensions. Among 25 interaction topics representing 25 million cell pairs, seven topics (2, 4, 7, 10, 18, 22, and 24) represented 55% of the total cell pair interactions, with each topic containing 3%+ cell pairs (Figure 2C). The other 18 interaction topics, each with 2% of the total cell pair interactions embedded baseline interaction signal. The most represented interaction topic was topic 22, consisting of 12% of the total cell pair interactions.

The model estimated the LR gene loadings in each interaction topic that described the relative contribution of each gene. These loadings can be ranked to identify biologically interpretable topic-specific top genes in each interaction topic (Figure 2D). Topics 22 and 24 captured immune-related interactions. Topic 22 was labelled as a lymphoid topic because top receptor genes included the known subunit of T-Cell Receptor Complex *CD3D* and killer cell lectin-like receptors *KLRC1*, *KLRC2*, and *KLRD1*. The top ligands in this topic are *HLA-E*, *CLEC2B*, and *CLEC2DC*, which are essential known modulators in cytotoxic T cells [Dufva et al., 2020]. Similarly, topic 24 was labelled as a myeloid topic as top genes in this topic showed enrichment of LR genes expressed by myeloid progenitors, for example - receptors such as *CD68*, *TREM2*, and *CR1*, and ligands such as *CCL23*, *CCL18*, *CCL13*, and *C1QA* [Hussain et al., 2021].

Topics 10 and 7 represented many oncogenes mutated in cancer. For instance, Topic 10 was cancer-growth associated, and genes that play a role in cancer cell survival and growth are enriched in this topic. The top receptors in this topic are growth factor receptors such as *ERBB2*, cell proliferation and growth signalling receptor *FZD10*, and immune inhibiting signalling receptor *ADORA2A* [Miller et al., 2015, Cekic and Linden, 2014]. Similarly, Topic 7 was a cancer-metastasis topic and genes such as *NTRK3*, known to increase the metastatic potential of cancer cells, *GRPR*,

which promotes EMT, and *UNC5A*, a known regulator of cancer plasticity, are enriched in this topic [Zhang et al., 2021b, Elshafae et al., 2016, Padua et al., 2018].

Moreover, Topic 18 was stroma-specific and represented genes that play an integral role in regulating the extracellular matrix (ECM) of the tumour immune microenvironment. These genes are highly expressed in cancer-associated fibroblast (CAF) and perivascular-like (PVL) cells. The top ligands in these topics are *COL1A1*, *COL1A2*, *COL3A1*, and *MMP13*, and the top receptors are *ITGA11* and *SCARA5* [Primac et al., 2019, bansal2017integrin]. Similarly, Topic 2 is enriched with genes highly expressed in endothelial cells, thus labeled an endothelial topic. Here, ligand proteins highly expressed in endothelial cells, such as *CD34*, *ANGPT2*, and *NID2* and receptors such as *APLNR* and *ESAM* are enriched [Wu et al., 2017]. Likewise, Topic 4 enriches genes involved in TME regulation processes and is thought to facilitate cancer progression and growth. The top genes in this topic include *KISS1R*/*KISS1*, which play a complex role in both restricting and promoting cancer cell survival, *IL20RB*, which promotes immunosuppressive microenvironment, and *MMP24*, which negatively regulates the aggressiveness of cancer cells [Cvetković et al., 2013].

The top LR genes in the major interaction topics show enrichment of different cell type-specific functional interactions. To confirm that each interaction topic captured cell type-specific CCC, we took a closer look at the distribution of cell types of target cells in each interaction topic for all the cells in the dataset. We found that the functional role of enriched top LR genes in each interaction topic matched with the dominant cell type of target cells in that topic (Figure 2E). For example, in the cancer-growth associated, on average, 68% of target cells for all cell types were cancer cells. Similarly, 49% of target cells in the stroma topic were CAF/PVL cells, and myeloid and T cells comprised 49% and 38% of target cells in the myeloid and lymphoid topics, respectively. For the endothelial topic, the dominant target cell type stemmed from endothelial cells, comprising 18%. In contrast, for the TME-regulation topic, both epithelial and plasms cells were dominant cells consisting of 20% and 22%, respectively.

In addition, the cell type-specific enrichment of interaction topics was further corroborated by the distribution of interaction topics in each cell type. Cancer cells, along with epithelial, plasma, and B cells, showed heterogeneous interaction patterns compared to myeloid, T, endothelial, CAF, and PVL cell types (Figure 2F). For cancer cells, the majority of interactions belonged to the cancer-growth and cancer-metastasis topics, where many of the top genes were oncogenes. Here, 55% of the total interactions were incident with a cancer-growth process, and 17% belonged to the cancer-metastasis topic, while the other five remaining topics consisted of 3-8% of interactions. Among the non-malignant cell types, the dominant interaction topic for myeloid cells was the myeloid interaction topic, and for T-cells, it was the lymphoid interaction topic. Here, 88% of myeloid cell and 65% of T cell interactions were found to be in respective interaction topics. Similarly, 83% of cell interactions with endothelial cells belonged to the endothelial topic, and 77% and 53% of interactions with CAF and PVL cells, respectively, were the stroma topic. In contrast, plasma, B, and epithelial cells showed a higher mixture of non-immune associated interaction topics.

Interaction topics provide a way to understand breast cancer heterogeneity.

Next, we investigated the heterogeneity of breast cancer cells based on the unbiased transcriptomic signature captured by the cell topic model while relating the cell topics to the interaction topics characterized by the SPRUCE analysis. The interaction patterns of 25,835 cancer cells manifested all the patterns of interactions (Figure 3B). The cell topic model identified 13 cell topics for cancer cells that show a distinct pattern of interactions with their target cells (Figure 3C). As expected,

the cancer growth topic was the most dominant (>57%) among 7 of 13 cell topics. For instance, 75%, 70%, and 68% of interactions for cancer cells in cell topics 24, 48, and 2 were driven by cancer-growth-related interactions. However, two cell topics were found more frequently interacting with non-cancer cells. The 66% of interactions involving (cell) Topic 9 were of stroma edges, and 60% of edges emanating from Topic 34 were assigned to endothelial interactions. Such a striking interaction heterogeneity was rarely observed in other cell types, such as T-cells, myeloid, endothelial, CAF, and PVL cell types, as they mostly interact within the same cell types. B-cells and epithelial cells, however, showed substantial variability of interaction patterns across cell topics.

Breast cancer cells are classified into subtypes based on the genomics and pathology of the disease, and different subtypes often result in markedly different clinical outcomes [Horr and Buechler, 2021]. All three different subtypes of cancer cells indeed exhibit diverse interaction patterns where TNBC (triple-negative breast cancer) cells were more heterogeneous compared to HER2+ and ER+ subtypes (Figure 3D). Here, more than 85% of interactions of the HER2+ subtype consisted of cancer-related topics, such as 82% for the cancer-growth and 5% for cancer metastasis topics. Similarly, for the ER+ subtype, more than 80% of interactions were cancer-related: 61% for the cancer-growth and 20% for the cancer-metastasis topics. In contrast, TNBC subtype cells were more diverse in interactions, with 42% for cancer growth, 15% for the cancer-metastasis, 15% for stroma, and 10% for myeloid topics.

Additionally, our approach identified a specific group of cells (cell topics) within these cancer subtypes that show topic-specific interaction patterns. For instance, TNBC cell topics show higher heterogeneity in interaction patterns at the cell topic level compared to the ER+ and HER2+ subtypes. The distribution of interaction patterns among cell topics is correlated with the expression pattern of LR genes enriched in each interaction topic. For example, TNBC cancer cells in cell topic 9 show higher expression of LR genes enriched in the myeloid interaction topic. In contrast, the cancer-growth LR genes are dominant among TNBC cancer cells in cell topic 24 (Figure 3E).

Interaction topics uncover underlying cancer-subtype-specific gene-gene interaction networks

Our approach also reveals topic-specific interaction patterns in the model parameter matrix (Fig. 4), with which gene-gene correlation networks can be estimated (ligand vs. receptor). For instance, gene networks in the cancer metastasis topic (Fig. 4A, burgundy strips; Fig. 4B, the first panel) consisted of a group of structural genes *CLDN4*, *LSR*, and *DSG2* involved in cell transformation and migration and signalling pathways *GPR37*, *CD151*, and *CD63* that are active in proliferation and migration, including epithelial-mesenchymal transition [Shang et al., 2012, Wang et al., 2021]. In the cancer growth topic (Fig 4A, green strips; Fig. 4B, the second panel), the resulting gene network primarily consisted of known oncogenes, such as *PTPRF*, *FGFR1*, *ERBB2*, and *TNFRSF1A* [Butti et al., 2018] and developmental genes, such as *LAMP1*, *ITGB1*, *RPSA*, *CANX*, *ATP6AP2*, and *MCFD2* [Going et al., 2018]. It is worth noting that the interaction networks adjacent to cancer cells generally include known cancer-related genes and other genes involved in cellular developmental process. Our analysis put them together in the same network modules, implicating a potential role of these interactions for cancer cells to hijack a normal cellular process. Similarly, immune-modulatory receptors primarily expressed in myeloid lineage cells (Fig. 4A, dark green; Fig 4B, the third panel) *TREM2*, *CSF1R*, *CSF2R*, *LILR*, and *IL3R* and signalling pathways *LTBR* and *TYROBP* required for the activation of myeloid cells are associated with myeloid-associated interaction topic. This topic captures the interactions between cancer cells and myeloid cell progenitors such as tumour-associated macrophages (TAM) in the tumour microenvironment, suppressing T cells and

facilitating tumour growth [Molgora et al., 2020].

In the lymphoid topic (Fig. 4B, the fourth panel), T cell receptor (TCR) complex (*CD3D*, *CD3G*, *CD2*, and *CD247*) and TCR signalling pathway genes (*PTPRC*, *CD45*, and *CD53*) are highly expressed both in cancer and surrounding immune cells [Shah et al., 2021]. Other chemokine receptor genes, such as *CXCR3* and *CXCR4*, were also found highly co-activated in this interaction topic, corroborating the pivotal role of crosstalk between T cells and cancer cells in promoting cancer growth, immune evasion, and metastasis [Kuo et al., 2018]. Interestingly, other killer-cell lectin-like receptors, which also co-occurred in this module, namely *KLRC1*, *KLRD1*, and *KLRF1*, are known to restrict T-cell's antitumour immunity [Hu et al., 2021]. The stroma topic (Fig. 4A, brown strips; Fig. 4B, the first panel of the second row) represented gene networks that capture the interaction of cancer cells with surrounding cells that promote its vascularization. It consisted of *NOTCH3*, *AVPR1A*, *MYLK* and integrin-mediated *ITGA1*, *ITGA5*, *ITGA7*, and *ITGB1* signalling pathways that play vital roles in tumour cell adhesion and progression [Price et al., 2020]. The genes *MCAM*, *ENPEP*, *EDNRA*, and *DCBLD2* that promote blood vessel formation and enhance tumorigenesis are enriched in this topic [Wragg et al., 2016]. Similarly, genes enriched in the endothelial topic recapitulate interactions of cancer cells in developing tumour vascular networks, especially in conjunction with endothelial cells. Previously known that *PECAM1*, *CALCR*, *ADGRL4*, and *CD93* genes are predominantly expressed in endothelial cells and regulate angiogenesis in tumour cells [Sheldon et al., 2021]. Additionally, TME-regulation associated topic primarily consisted of genes mixture of endothelial-associated and stroma-associated topics with enrichment of distinct genes known to control tumour growth and promote stemness of cancer cells in microenvironment such as *KCNN4*, *IL6ST*, and *CD1B* [Fan et al., 2022].

Discussion

No single cell can exist alone in human tissues. We propose a novel machine learning framework that systematically dissects tens of millions of cell-cell pairs and uncovers common patterns of how cells talk to each other concerning cell surface ligand-receptor protein interactions. In particular, our analysis focused on finding commonly used communication channels in breast cancer progression and metastasis by reanalyzing state-of-the-art single-cell genomics data sets. Our approach, built on probabilistic topic modelling and variational autoencoder model, specifically demonstrated that a part of cancer heterogeneity could be understood in diverse and context-specific interaction partners of cancer cells. We found that many ligand-receptor interactions can occur in a subtype-specific manner, although cells are largely clustered as cancer cells in conventional single-cell analysis. Along the line, our results suggest that cell types and states are better understood, and the definitions of cell types can be refined while considering cell-cell communication patterns.

Our proposed approach generalizes existing bioinformatics methods and does not rely on prescribed cell-type annotations/clustering results, which may introduce unwanted biases in downstream analysis. Moreover, if cells within a cluster are not homogeneous as anticipated, a clustering-based cell-cell commutation method can easily result in confounded correlation statistics, clearly violating necessary assumptions, such as independent and identification distributed expression values. Here, we differently formulate cell-cell communications analysis from an edge's perspective, whereby a single edge (interaction) is a data point, and the feature vector can be engineered by exploiting the information of both endpoints of the edge (ligand and rector expression values). Such a novel formulation is better suited for the analysis of large-scale single-cell data and also easily extends to principled data integration strategies. For instance, if cell-cell interaction pairs were already con-

structed by spatial transcriptomics data, we can easily construct feature vectors by combining two gene expression vectors (one from the source and the other from target cells). For the multiomics data integration tasks, we can concatenate multiple data modalities to investigate the co-occurrence of multimodal expressions, such as DNA accessibility, histone modifications, and metabolomics.

We acknowledge that our SPRUCE approach relies on several specialized modelling assumptions. One of which is that we assume that known ligand-receptor protein-protein interaction networks serve as a superset/backbone of topic-specific interaction networks. Considering that most protein-protein interactions were experimentally discovered *in vitro* by error-prone high-throughput methods, there is room for improvements in terms of the precision and specificity of the interaction analysis method. Here, we only focused on immediate surface protein interactions. However, establishing causal effects on the downstream genes emanating from surface protein singling pathways can further enrich our understanding of disease etiology.

Acknowledgments

This work was supported by the BC Cancer Foundation and NSERC Discovery Grant (YPP). We acknowledge financial support from UBC Four Year Fellowship (SS).

Author contributions

SS and YPP conceived the idea. SS implemented variational autoencoder models. SS also wrote analysis scripts. SS and YPP wrote this manuscript.

Declaration of interests

Nothing to declare.

Figure legends

Figure 1. Probabilistic topic model identifies cell topics for resident cell types and cancer subtypes. (A) Distribution of cell types highlighting the total number of cells in each cell topic. Relative proportion of cell types in each cell topic. Cell topic is assigned to each cell based on the highest score. (B) UMAP visualization of 50 cell topics representing an integrated dataset consisting of 155,913 cells. Each dot is a single cell, and colors represent the corresponding cell type clustered according to k-means(n=50) algorithm on cell topic proportions and annotated using the majority voting rule with annotation from previous studies and SingleR. (C) Heatmap of top 25 gene loadings associated with each cell topic with the total number of cells > 100. Cell topics (y axis) and genes (x axis) are ordered according to hierarchical clustering (optimal leaf ordering). The cell type and topic are shown on the right y-axis, while the top 3 genes for each cell topic are depicted on the left y-axis. (D) Log normalized (total count to 10,000 reads per cell) expression of markers genes for cell types- *EPCAM* for epithelial, *CD3D* for T-cells, *CD68* for myeloid cells, *MS4A1* for B-cells, *PECAM1* for endothelial cells, and *PDGFRB* for CAF/PVL cells. (E) Relative proportion of 50 cell topics from a sample(n=50) of cells assigned to cell type - cell topic pairs.

Figure 2. SPRUCE model overview and the common interaction patterns idenfied by the model. (A) UMAP-based representation of cell pairing method where source cell is paired with four different target cells from each cell topic. (B) Given a cell pair LR gene expression data as input, SPRUCE model transforms and aggregates interaction-driven LR data space and feeds into the neural network. The encoder learns the latent topic representation of the interaction between cells using the mixture of experts approach from ligand and receptor encoders. The decoder generates biologically interpretable topic embeddings separately over the ligand and receptor genes. (C) The distribution of ~25 million cell pair interactions over 25 interaction topics shows seven major interaction patterns. (D) Heatmap of top 10 gene loadings associated with seven interaction topics. Interaction topics (x axis) and LR genes (y axis) are ordered according to hierarchical clustering (optimal leaf ordering). (E) Relative proportion of target cell type in all the source cell types associated with the major interaction patterns. Each row is source cell type (y axis) and target cell type proportions (x-axis). (F) Enrichment of interaction patterns among cell types.

Figure 3. Heterogeneity of cancer cells revealed by the cell-topic specific interaction patterns. (A) Representation of interaction topics of cancer cells with surrounding cell types in the TME. (B) Structure plot showing the variability of interaction topic proportion estimates among 25,835 cancer cells in the dataset. Proportion of seven interaction topics (y axis) and cancer cell pair (x-axis) with one representative target cell among 159 source-target cell pairs are clustered using k-means(n=7) clustering algorithm using interaction topic proportions. (C) The proportion of interaction topics of cancer and non-malignant cells associated with cell topics. Boxplots show the distribution of interaction topic proportions for each interaction topic across all cell topics. (D) The proportion of interaction topics of cancer subtypes associated with cell topics. Boxplots depict the estimated interaction topic proportions across all cell topics. (E) Log normalized (total count to 10,000 reads per cell) expression of LR genes from all the cells associated with respective cancer subtype and cell topic. The genes (y axis) are top 25 LR genes from seven interaction topics with expression values > 0.05.

Figure 4. Gene network derived from the interaction topics. (A) Circular chord diagram of the interaction topic-gene network representing significantly (Z-score > 4.0) enriched LR gene loadings in each interaction topic. The edge between LR genes and interaction topics shows the occurrence of a gene in LR pairing and if a gene is unique to or common among different interaction topics.

The genes with (*) symbol are among the top 10 LR genes based on loadings estimated by the interaction topic model. (B) Ligand-receptor bipartite network for each interaction topics depicting significantly (Z-score > 4.0) enriched LR genes. The edge in the graphs represents the magnitude of Pearson correlation coefficients. The top 100 LR edges in each interaction topic with a correlation coefficient > 3.0 are selected.

Tables

STAR Methods

Data preprocessing

We constructed a dataset to represent an immune-enriched breast cancer microenvironment by combining cancer cells with immune cells and healthy cells from three recent breast cancer-related studies. The breast cancer dataset consists of 100k cells from a single-cell atlas of human breast cancers [Wu et al., 2021]. The source of the normal dataset consisting 48k cells is a single-cell atlas of the healthy breast tissues [Bhat-Nakshatri et al., 2021]. The third dataset composed of immune cells is 6k subset of breast cancer CD4 and CD8 T-cells from a pan-cancer atlas of tumour-infiltrating T cells profiled across 21 cancer types and 316 donors [Zheng et al., 2021]. The total number of cells in the combined dataset is 155,913. We filtered out genes detected in less than three cells along with mitochondrial and spike genes, leading to 20,265 genes in the final dataset.

Cell-level probabilistic topic modelling

SPRUCE utilizes composite embedded topic models (ETM) to identify cell subtypes/states and define its interaction pattern based on cell-cell communication in the tumour microenvironment (TME). The input for the model is single-cell data, which contains multiple cell types present in the TME across different data sets. Let X_{iG} be raw count gene expression data for G genes in cell i. We modelled X_{iG} with multinomial distribution parameterized by a normalized gene expression frequency, achieving a scale-invariant property across different cells, batches, and data sets. Further, we introduce Dirichlet prior on the gene frequency and parameterize the Dirichlet as a generalized linear model(GLM) with linear combinations of topic-specific probabilities.

The model consists of two ETMs, each with a pair of encoder-decoder networks. The first ETM, the cell topic model, takes gene expression single-cell data as input and models cell topic and topic-specific gene loadings. Next, we assigned a cell topic to each cell based on the highest topic proportion from the cell topic model. The topic assignment was used to construct a set of target cells for each cell such that one cell is paired with the five nearest target cells from each topic. The LR gene expression data from each cell pair were transformed into each other's space by a binary cell-interaction database- celltalkdb. The transformed ligand and receptor data were treated as two independent modules with separate encoder and decoder modules in the model. The latent variables with encoded information from two modules were combined by taking their average to obtain a final interaction topic variables as a mixture of experts from the ligand and receptor latent space. The second ETM, the interaction topic model, uses transformed LR data from source-target cell pairs as input and models interaction topic for each cell pair with topic-specific LR gene loadings.

Variational autoencoder model for topic modelling

The cell topic model extends on the ideas of LDA where each cell is considered as a document and each gene in the cell as a word. Here, a sample of cells $i_1, \dots, i_N$ and a list of genes $g_1, \dots, g_G$, where $x_{i1}, x_{i2}, \dots, x_{iG}$ are raw count data for G genes in cell i. We assume that each cell i is represented as a mixture of latent cell topics with topic proportion θ_{it}. θ_{it} is drawn from logistic normal distribution with hyperparameters δ_{it}, which are learnable parameters of a neural network. The cell topic proportions are within a simplex, namely $\sum_{t \in \mathsf{topics}} \theta_{it} = 1$.

$$\delta_{it} \sim \mathcal{N}(0, I); \quad \theta_{it} = h(\delta_{it}),$$

where $h(\mathbf{x}) = \exp(x_{it}) / \sum_{t'} \exp(x_{it'})$.

We assume that each cell count vector X_i was generated from a multinomial distribution parameterized by a gene expression frequency matrix with an element ρ_{ig} of a gene g in the cell i.

$$p(\mathbf{X}_i|\rho_i) = \prod_g \rho_{ig}^{x_{ig}}$$

In conventional ETM, ρ is directly modelled by transforming each cell's topic proportion θ_{it} in a topic space to a gene space as a linear combination of topic-specific probabilities, $\rho_{ig} = \sum_t \theta_{it}\beta_{tg}$, where β_{tg} captures a topic t specific frequency of a gene g. Instead of modelling ρ directly, we introduced Dirichlet prior on the ρ and parameterize the Dirichlet as a generalized linear model (GLM). Exploiting the conjugacy between the multinomial and Dirichlet, we integrate out the unknown parameters ρ. Further, we draw β_{tg} from normal distribution.

$$p(\rho_i|\lambda_i) = \frac{\Gamma(\sum_g \lambda_{ig})}{\prod_g \Gamma(\lambda_{ig})} \prod_{g \in \text{genes}} \rho_{ig}^{\lambda_{ig}-1}$$

$$\lambda_{ig} = \exp\left(\sum_{t=1}^{T} \theta_{it}(\beta_{tg} + b_g)\right)$$

$$\beta_{tg} \sim \mathcal{N}(0, I)$$

The marginal likelihood of a single cell data X_i under a generative process is described as:

$$p(\mathbf{X}_i|\cdot) = \frac{\Gamma(\sum_g \lambda_{ig})}{\sum_g \Gamma(\lambda_{ig})} \frac{\Gamma(\sum_g \lambda_{ig} + x_{ig})}{\sum_g \Gamma(\lambda_{ig} + x_{ig})},$$

where $\Gamma(\cdot)$ is the Euler's gamma function.

The posterior distribution $p(\delta_i, \beta_{gi}|X_i)$ is intractable because it involves integral over the cell topic proportion. Variational inference techniques can be used to approximate this type of intractable integrals. Here, we use mean-field variational distribution $q(\delta_i, \beta_{gi}|X_i) = q(\delta_i|X_i)q(\beta_{gi}|X_i)$ to approximate the true posterior. Both distributions are chosen to be Gaussian where mean and variance for $q(\delta_i|X_i)$ and $(\beta_{gi}|X_i)$ distributions are generated by the encoder and the decoder networks, respectively.

To minimize the Kullback-Leibler (KL) divergence between the true posterior and the approximate posterior, we maximize the evidence lower bound (ELBO) of the log-likelihood:

$$ELBO = E_{q(\delta|X)}[\log\ p(X \mid \delta)]$$

$$-D_{KL}(q(\delta|X)\|p(\delta|X))$$

$$-D_{KL}(q(\beta|X)\|p(\beta|X))$$

The encoder for the cell topic model consisted of a 4-layered neural network with two hidden layers of size 200 and two with sizes 100 and 50. We used an Adam optimizer with a learning rate 0.01 and optimized the model for convergence for 1000 epochs with a minibatch size of 128.

Cell type labelling by propagating within topic clusters

To assign each cell to a cell topic, we applied our proposed cell-topic model to an integrated dataset representing the TME of breast cancer, which consisted of 155,913 cells and 20,265 genes. After unsupervised training of the model, all cells were mapped in the latent cell topic space. Clustering on latent variables was performed using k-means, and clusters were mapped to known cell populations using the majority rule. Annotations for cell type were assigned to each cell based on an assignment from previous studies. In cases with missing annotation, we used the reference-based cell type identification method SingleR with a tumour microenvironment reference dataset from CHETAH [Aran et al., 2019, De Kanter et al., 2019].

Cell-cell interaction topic modelling

Construction of cell-cell interaction networks

A target cell set was constructed for each cell using the topic assignment from the cell topic model. For each cell, the five closest target cells from each topic were calculated using python package [ANN] with angular distance on the cell topic space. An annoy model was created for each topic with a total cell count greater than 100. In total, we generated 159 target cell pairs for each cell - 32 topics and 5 target cells from each topic, excluding a self-target cell.

We transformed the raw LR gene expression data into each other's space for each cell pair using a binary interaction matrix generated from a publicly available database - celltalkDB [Shao et al., 2021]. Here, $A_{lr} = 1$ if and only if a ligand l binds with a receptor protein r in the cell interaction database; otherwise, $A_{lr} = 0$.

For a cell pair $e \equiv (i, j)$, transformed count data $Y_{e_{ij}g}$ for gene g is defined as:

$$Y_{e_{ij}g} = (i_l \times A_{lr} \times i_r) + (j_l \times A_{lr} \times j_r),\ g \in (L, R)$$

Interaction topic model

The interaction topic model uses transformed ligand and receptor (LR) gene expression data from source-target cell pairs to model the interaction topic for each cell pair and identify LR genes enriched in each interaction topic. The model follows the architecture of the cell topic model, where each cell pair is considered a document and LR genes in the cell pair as words. Here, a sample of cell pair $e_1, \dots, e_N$ and a list of ligand genes $l_1, \dots, l_L$ and receptor genes $r_1, \dots, r_R$, where $y_{el_1}, y_{el_2}, \dots, y_{eL}$ and $y_{er_1}, y_{er_2}, \dots, y_{eR}$ are transformed count data for L ligand and R receptor genes. We assume that each cell pair e is represented as a mixture of latent interaction topics with topic proportion θ_{et} generated from two gene modules of LR genes. We draw θ_{et} from logistic normal distribution with hyperparameters δ_{et} which is generated using a mixture of experts approach from two gene modules:

$$\delta_{et_l} \sim \mathcal{N}(0, I); \delta_{et_r} \sim \mathcal{N}(0, I)$$

$$\theta_{et} = h(\delta_{et}); \quad \delta_{et} = (\delta_{et_l} + \delta_{et_r})/2,$$

where $h(\mathbf{x}) = \exp(x_{it})/sum_{t'} x_{it'}$.

We assume that for each cell pair e transformed LR count vector Y_{eL} and Y_{eR} are generated from a multinomial distribution parameterized by a ligand expression frequency matrix with an element ρ_{el} for ligand l and a receptor expression frequency matrix with an element ρ_{er} for receptor r.

$$\begin{aligned} p(Y_e|\rho_{el}, \rho_{er}) &= p(\mathbf{Y}_{el}|\rho_{el})p(\mathbf{Y}_{er}|\rho_{er}) \\ &= \prod_l \rho_{el}^{y_{el}} \prod_r \rho_{er}^{y_{er}} \end{aligned}$$

Similar to the cell topic model, we added Dirichlet prior on the both ρ's and parameterize the Dirichlet as a generalized linear model (GLM) with linear combinations of topic proportion θ_{et} and topic-specific LR gene expression frequency β_{tg}. Next, we integrate out the unknown parameter ρ using the conjugacy between multinomial and Dirichlet distributions. Additionally, we draw β_{tl} and β_{tr} from normal distribution.

$$\rho_{eg} \sim \mathsf{Dir}(\rho_{eg}|\lambda_{eg}) = \frac{\Gamma(\sum_g \lambda_{eg})}{\prod_g \Gamma(\lambda_{eg})} \prod_g \rho_{eg}^{\lambda_{eg}-1}$$

$$\lambda_{eg} = \lambda_0 \exp\left(\sum_{t=1}^{T} \theta_{et}(\beta_{tg} + \delta_g)\right)$$

$$\lambda_0 = \exp(\tilde{\lambda}_0); g \in (L, R)$$

$$\beta_{tl} \sim \mathcal{N}(0, I); \beta_{tr} \sim \mathcal{N}(0, I)$$

The marginal likelihood of transformed cell pair data y_e under a generative process is described as:

$$p(\mathbf{Y}_e|\cdot) = p(\mathbf{Y}_{eL}|\cdot)p(\mathbf{Y}_{eR}|\cdot)$$

$$p(\mathbf{Y}_e|\cdot) = \frac{\Gamma(\sum_g \lambda_{eg})}{\sum_g \Gamma(\lambda_{eg})} \frac{\Gamma(\sum_g \lambda_{eg} + y_{eg})}{\sum_g \Gamma(\lambda_{eg} + y_{eg})}$$

$$g \in (L, R)$$

We used variational distribution to approximate the true posterior distribution as $p(\delta_{et}, \beta_{tl}, \beta_{tr}, |Y_e) = q(\delta_{et_l}|Y_{e_l})q(\delta_{et_r}|Y_{e_r})q(\beta_{tl}|Y_{e_l})q(\beta_{tr}|Y_{e_r})$. All approximate distributions are chosen to be Gaussian where mean and variance for $q(\delta_{et_l}|Y_{e_l})$ and $q(\delta_{et_r}|Y_{e_r})$ distrbutions are generated by the encoder, and $q(\beta_{tl}|Y_{e_l})$ and $q(\beta_{tr}|Y_{e_r})$ distributions are generated by the decoder networks.

To minimize the Kullback-Leibler (KL) divergence between the true posterior and the approximate posterior, we maximize the evidence lower bound (ELBO) of the log-likelihood:

$$
\begin{aligned}
\mathsf{ELBO} \;=\; & \mathbb{E}_{q(\delta_{et_l}|Y_{et_l})}\left[\log\ p(Y_{et_l}|\delta_{et_l})\right] \\
& + \mathbb{E}_{q(\delta_{et_r}|Y_{et_r})}\left[\log\ p(Y_{et_r}|\delta_{et_r})\right] \\
& - \mathsf{D}_{\mathsf{KL}}\left[q(\delta_{et_l}|Y_{el})\middle\|p(\delta_{et_l}|Y_{el})\right] \\
& - \mathsf{D}_{\mathsf{KL}}\left[q(\delta_{et_r}|Y_{er})\middle\|p(\delta_{et_r}|Y_{er})\right] \\
& - \mathsf{D}_{\mathsf{KL}}\left[q(\beta_{tl})\|p(\beta_{tl})\right] \\
& - \mathsf{D}_{\mathsf{KL}}\left[q(\beta_{tr})\|p(\beta_{tr})\right]
\end{aligned}
$$

The encoder for the cell topic model consisted of a 3-layered neural network with hidden layers of sizes 200, 100, and 25. We used an Adam optimizer with a learning rate of 0.01 and optimized the model for convergence for 500 epochs with a minibatch size of 5088 cell pairs (32 individual cells in a batch).

Topic-specific gene co-expression network

For each interaction topic t and LR gene g, we calculated a z-score s_{tg} as $\mathbb{E}\left[\beta_{tg}\right] / \sqrt{\mathbb{V}\left[\beta_{tg}\right]}$, where β_{tg} is topic-specific gene frequency estimated by the interaction topic model. The highest proportion estimate was used to assign an interaction topic to all cancer cell pairs. A gene co-expression network for each interaction topic was constructed using significantly enriched LR genes (z-score $>$ 4.0). The edge in the network represents the magnitude of Pearson correlation coefficient between LR gene pairs based on expression data in all the cancer cell pairs assigned to respective interaction topic.

References

ANNOY library. https://github.com/spotify/annoy. Accessed: 2022-03-01.

Axel A Almet, Zixuan Cang, Suoqin Jin, and Qing Nie. The landscape of cell–cell communication through single-cell transcriptomics. *Current opinion in systems biology*, 26:12–23, 2021.

Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2): 163–172, 2019.

Erick Armingol, Hratch M Baghdassarian, Cameron Martino, Araceli Perez-Lopez, Caitlin Aamodt, Rob Knight, and Nathan E Lewis. Context-aware deconvolution of cell–cell communication with tensor-cell2cell. *Nature communications*, 13(1):1–15, 2022.

Poornima Bhat-Nakshatri, Hongyu Gao, Liu Sheng, Patrick C McGuire, Xiaoling Xuei, Jun Wan, Yunlong Liu, Sandra K Althouse, Austyn Colter, George Sandusky, et al. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Reports Medicine*, 2(3):100219, 2021.

Ramesh Butti, Sumit Das, Vinoth Prasanna Gunasekaran, Amit Singh Yadav, Dhiraj Kumar, and Gopal C Kundu. Receptor tyrosine kinases (rtks) in breast cancer: signaling, therapeutic implications and challenges. *Molecular cancer*, 17(1):1–18, 2018.

Caglar Cekic and Joel Linden. Adenosine a2a receptors intrinsically regulate cd8+ t cells in the tumor microenvironmentadenosine maintains cd8+ t cells in solid tumors. *Cancer research*, 74 (24):7239–7249, 2014.

Donna Cvetković, Andy V Babwah, and Moshmi Bhattacharya. Kisspeptin/kiss1r system in breast cancer. *Journal of Cancer*, 4(8):653, 2013.

Jurrian K De Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank CP Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research*, 47(16):e95–e95, 2019.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Olli Dufva, Petri Pölönen, Oscar Brück, Mikko AI Keränen, Jay Klievink, Juha Mehtonen, Jani Huuhtanen, Ashwini Kumar, Disha Malani, Sanna Siitonen, et al. Immunogenomic landscape of hematological malignancies. *Cancer Cell*, 38(3):380–399, 2020.

Mirjana Efremova, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4):1484–1506, 2020.

Said M Elshafae, Bardes B Hassan, Wachiraphan Supsavhad, Wessel P Dirksen, Rachael Y Camiener, Haiming Ding, Michael F Tweedle, and Thomas J Rosol. Gastrin-releasing peptide receptor (grpr) promotes emt, growth, and invasion in canine prostate cancer. *The Prostate*, 76 (9):796–809, 2016.

Jing Fan, Ruofei Tian, Xiangmin Yang, Hao Wang, Ying Shi, Xinyu Fan, Jiajia Zhang, Yatong Chen, Kun Zhang, Zhinan Chen, et al. Kcnn4 promotes the stemness potentials of liver cancer

stem cells by enhancing glucose metabolism. *International journal of molecular sciences*, 23(13): 6958, 2022.

Catherine C Going, Dhanir Tailor, Vineet Kumar, Alisha M Birk, Mallesh Pandrala, Meghan A Rice, Tanya Stoyanova, Sanjay Malhotra, and Sharon J Pitteri. Quantitative proteomic profiling reveals key pathways in the anticancer action of methoxychalcone derivatives in triple negative breast cancer. *Journal of proteome research*, 17(10):3574–3585, 2018.

Christina Horr and Steven A Buechler. Breast cancer consensus subtypes: A system for subtyping breast cancer tumors based on gene expression. *NPJ breast cancer*, 7(1):1–13, 2021.

Ziming Hu, Xiuxiu Xu, and Haiming Wei. The adverse impact of tumor microenvironment on nk-cell. *Frontiers in Immunology*, 12:633361, 2021.

Khiyam Hussain, Mark S Cragg, and Stephen A Beers. Remodeling the tumor myeloid landscape to enhance antitumor antibody immunotherapies. *Cancers*, 13(19):4904, 2021.

Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1–20, 2021.

Paula T Kuo, Zhen Zeng, Nazhifah Salim, Stephen Mattarollo, James W Wells, and Graham R Leggatt. The role of cxcr3 and its chemokine ligands in skin disease and cancer. *Frontiers in Medicine*, 5:271, 2018.

Martin L Miller, Ed Reznik, Nicholas P Gauthier, Bülent Arman Aksoy, Anil Korkut, Jianjiong Gao, Giovanni Ciriello, Nikolaus Schultz, and Chris Sander. Pan-cancer analysis of mutation hotspots in protein domains. *Cell systems*, 1(3):197–209, 2015.

Martina Molgora, Ekaterina Esaulova, William Vermi, Jinchao Hou, Yun Chen, Jingqin Luo, Simone Brioschi, Mattia Bugatti, Andrea Salvatore Omodei, Biancamaria Ricci, et al. Trem2 modulation remodels the tumor myeloid landscape enhancing anti-pd-1 immunotherapy. *Cell*, 182(4):886–900, 2020.

Maria B Padua, Poornima Bhat-Nakshatri, Manjushree Anjanappa, Mayuri S Prasad, Yangyang Hao, Xi Rao, Sheng Liu, Jun Wan, Yunlong Liu, Kyle McElyea, et al. Dependence receptor unc5a restricts luminal to basal breast cancer plasticity and metastasis. *Breast Cancer Research*, 20(1):1–18, 2018.

Jessica C Price, Elham Azizi, LA Naiche, Jenny G Parvani, Priyanka Shukla, Seoyeon Kim, Jill K Slack-Davis, Dana Pe'er, and Jan K Kitajewski. Notch3 signaling promotes tumor cell adhesion and progression in a murine epithelial ovarian cancer model. *PloS one*, 15(6):e0233962, 2020.

Irina Primac, Erik Maquoi, Silvia Blacher, Ritva Heljasvaara, Jan Van Deun, Hilde YH Smeland, Annalisa Canale, Thomas Louis, Linda Stuhr, Nor Eddine Sounni, et al. Stromal integrin α11 regulates pdgfrβ signaling and promotes breast cancer progression. *The Journal of clinical investigation*, 129(11):4609–4628, 2019.

Kinjal Shah, Amr Al-Haidari, Jianmin Sun, and Julhash U Kazi. T cell receptor (tcr) signaling in health and disease. *Signal transduction and targeted therapy*, 6(1):1–26, 2021.

Xiying Shang, Xinjian Lin, Edwin Alvarez, Gerald Manorek, and Stephen B Howell. Tight junction proteins claudin-3 and claudin-4 control tumor growth and metastases. *Neoplasia*, 14(10):974–IN22, 2012.

Xin Shao, Jie Liao, Chengyu Li, Xiaoyan Lu, Junyun Cheng, and Xiaohui Fan. Celltalkdb: a manually curated database of ligand–receptor interactions in humans and mice. *Briefings in bioinformatics*, 22(4):bbaa269, 2021.

Helen Sheldon, Esther Bridges, Ildefonso Silva, Massimo Masiero, David M Favara, Dian Wang, Russell Leek, Cameron Snell, Ioannis Roxanis, Mira Kreuzer, et al. Adgrl4/eltd1 expression in breast cancer cells induces vascular normalization and immune suppressioneltd1 is angiogenic and immunosuppressive in breast cancer. *Molecular Cancer Research*, 19(11):1957–1969, 2021.

Keely Tan and Matthew J Naylor. Tumour microenvironment-immune cell interactions influencing breast cancer heterogeneity and disease progression. *Frontiers in Oncology*, 12, 2022.

Sarah Teichmann and Mirjana Efremova. Method of the year 2019: single-cell multimodal omics. *Nat. Methods*, 17(1):2020, 2020.

Koki Tsuyuzaki, Manabu Ishii, and Itoshi Nikaido. Uncovering hypergraphs of cell-cell interaction from single cell rna-sequencing data. *BioRxiv*, page 566182, 2019.

Jian Wang, Min Xu, Dan-Dan Li, Wujikenayi Abudukelimu, and Xiu-Hong Zhou. Gpr37 promotes the malignancy of lung adenocarcinoma via tgf-β/smad pathway. *Open Medicine*, 16(1):024–032, 2021.

Shuxiong Wang, Matthew Karikomi, Adam L MacLean, and Qing Nie. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic acids research*, 47(11):e66–e66, 2019.

Joseph W Wragg, Jonathan P Finnity, Jane A Anderson, Henry JM Ferguson, Emilio Porfiri, Rupesh I Bhatt, Paul G Murray, Victoria L Heath, and Roy Bicknell. Mcam and lama4 are highly enriched in tumor blood vessels of renal cell carcinoma and predict patient outcome. *Cancer research*, 76(8):2314–2326, 2016.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.

Xinqi Wu, Anita Giobbie-Hurder, Xiaoyun Liao, Courtney Connelly, Erin M Connolly, Jingjing Li, Michael P Manos, Donald Lawrence, David McDermott, Mariano Severgnini, et al. Angiopoietin-2 as a biomarker and target for immune checkpoint therapy. *Cancer immunology research*, 5(1): 17–28, 2017.

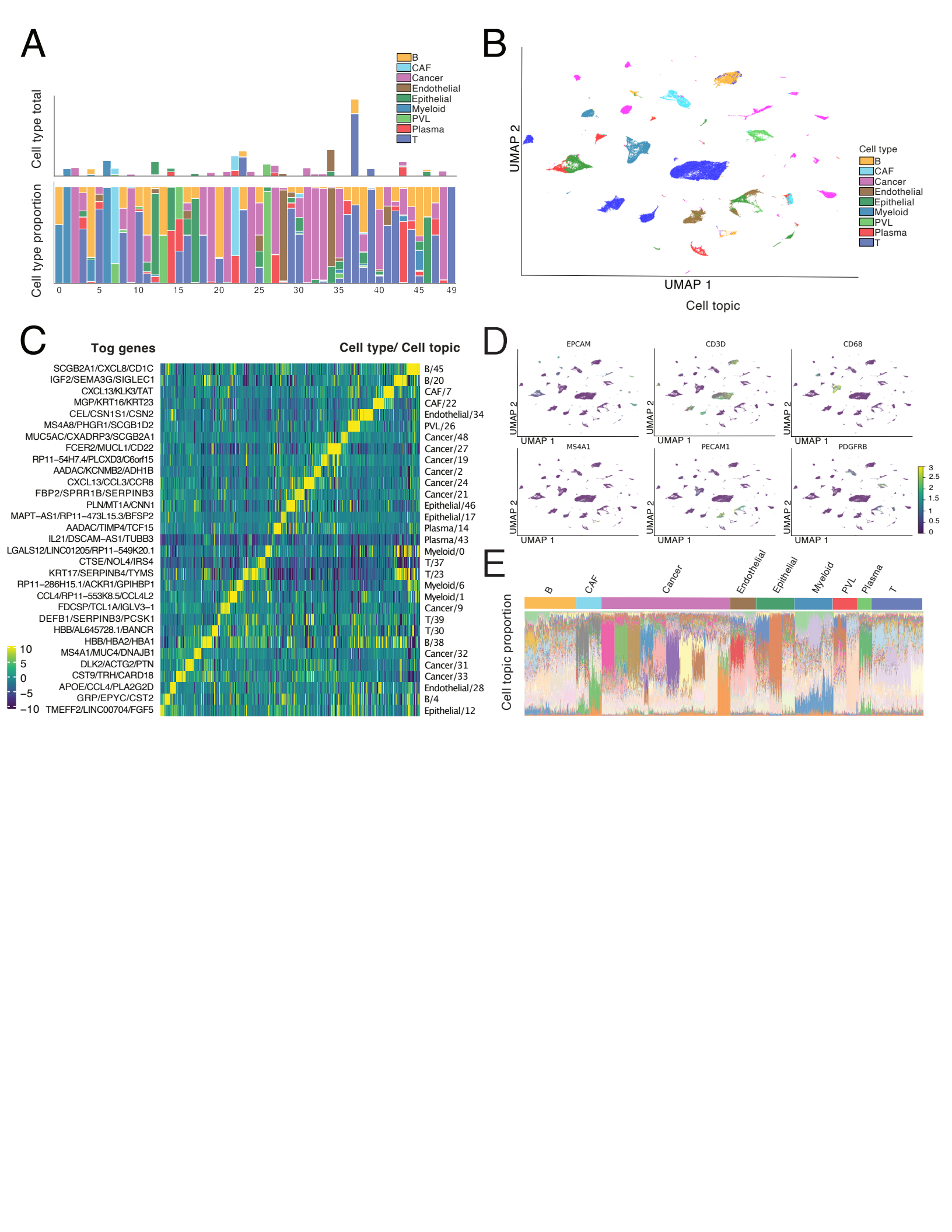
Yang Zhang, Tianyuan Liu, Xuesong Hu, Mei Wang, Jing Wang, Bohao Zou, Puwen Tan, Tianyu Cui, Yiying Dou, Lin Ning, et al. Cellcall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic Acids Research*, 49(15):8520–8534, 2021a.

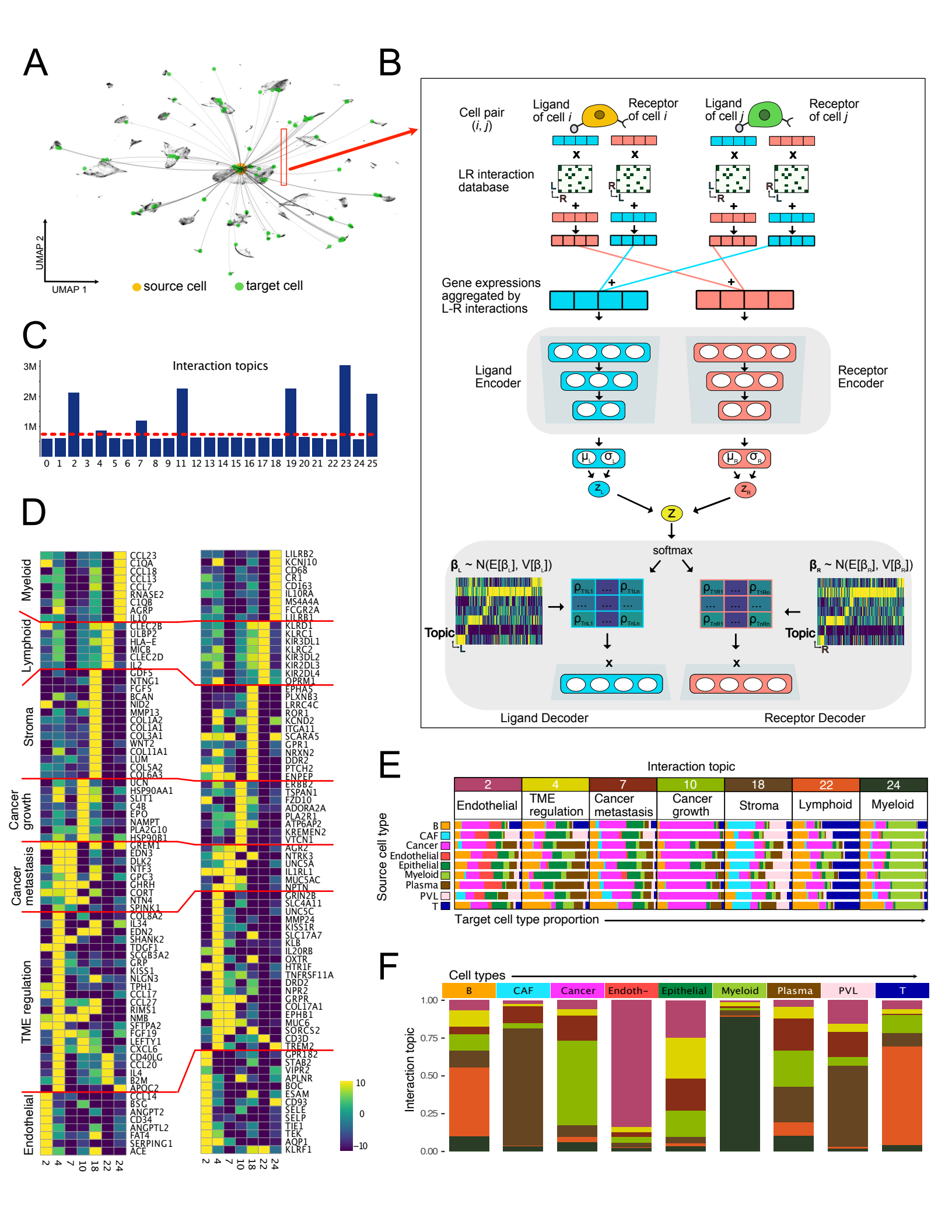
Zhao Zhang, Yongbo Yu, Pengfei Zhang, Guofeng Ma, Mingxin Zhang, Ye Liang, Wei Jiao, and Haitao Niu. Identification of ntrk3 as a potential prognostic biomarker associated with tumor mutation burden and immune infiltration in bladder cancer. *BMC cancer*, 21(1):1–13, 2021b.

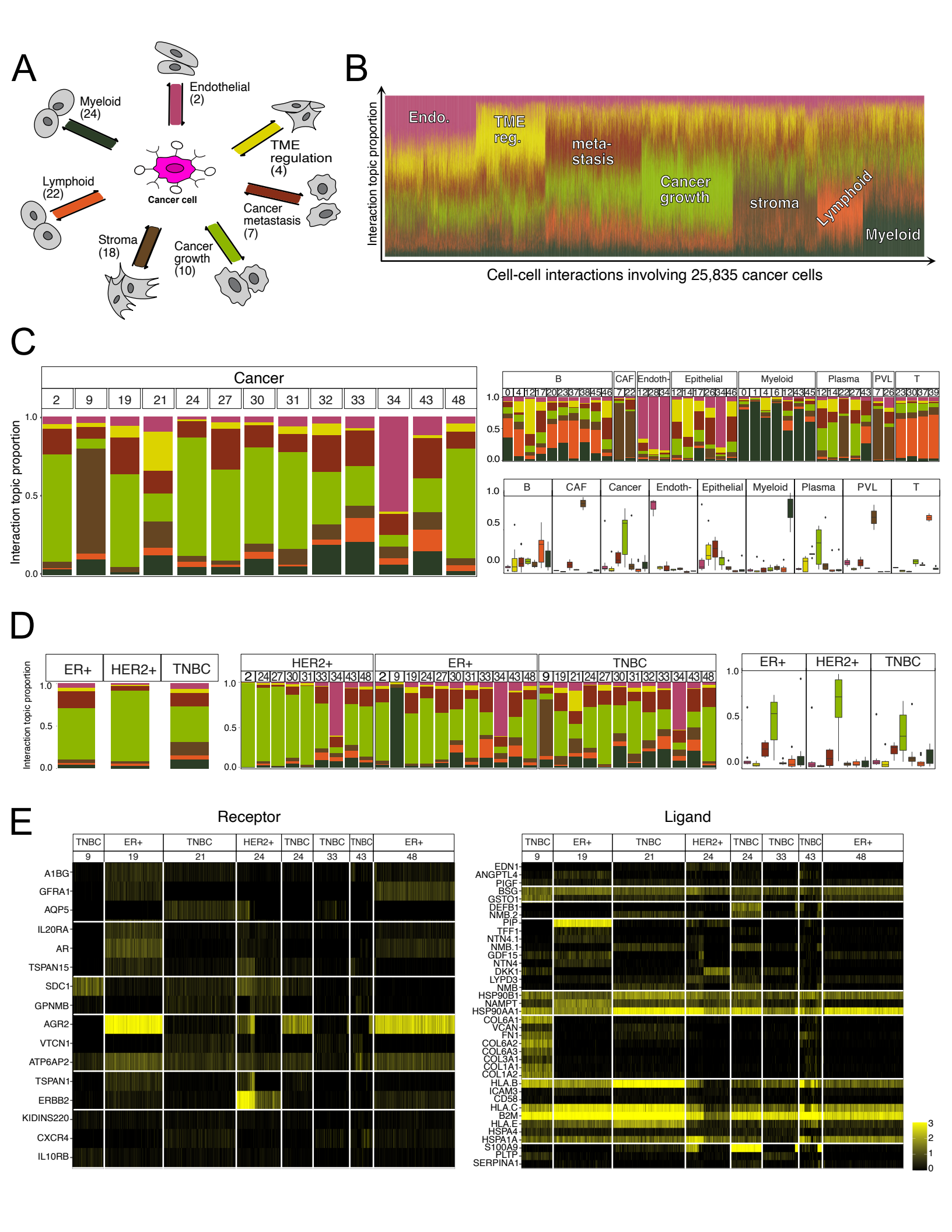
Yifan Zhao, Huiyu Cai, Zuobai Zhang, Jian Tang, and Yue Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications*, 12(1): 1–15, 2021.

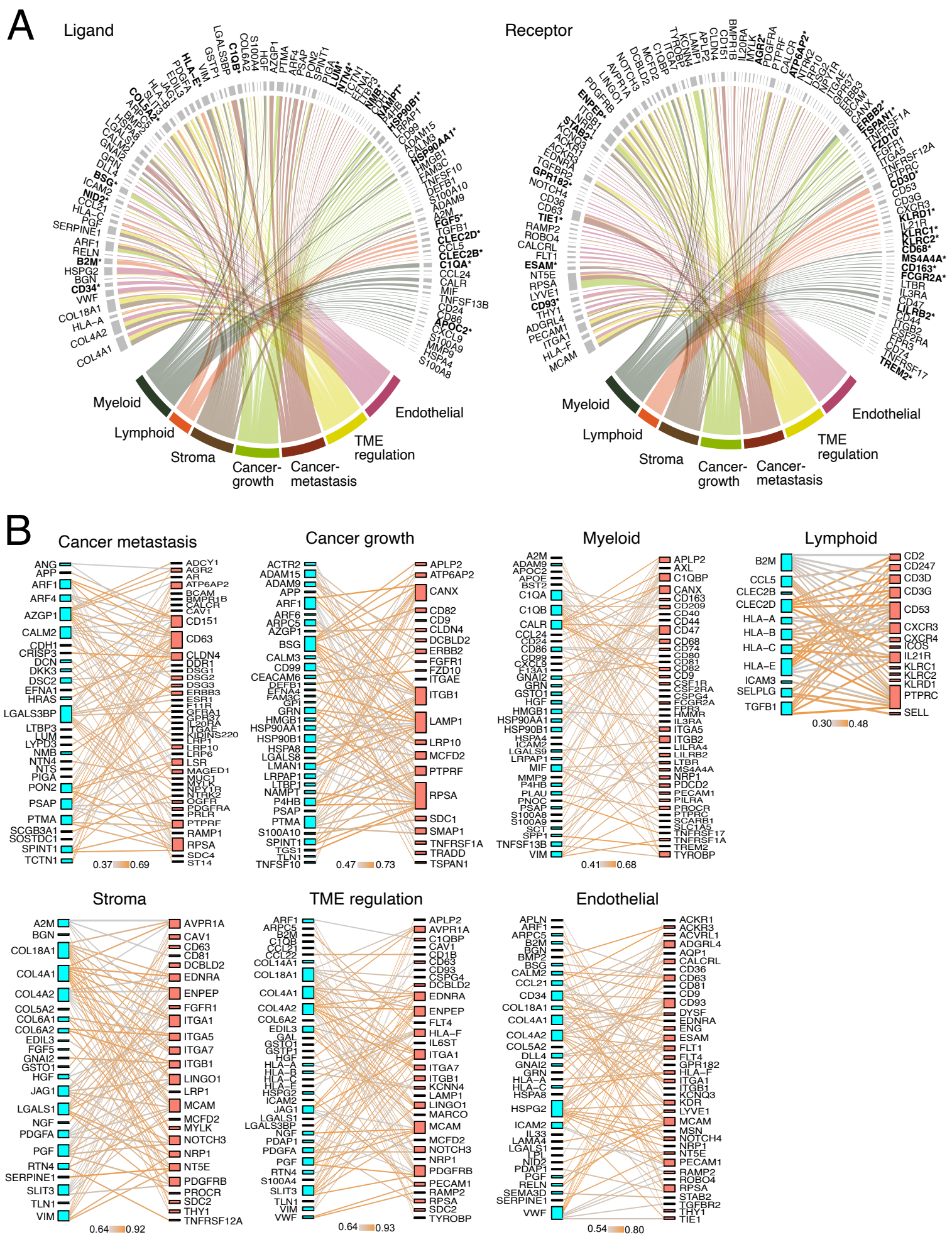
Liangtao Zheng, Shishang Qin, Wen Si, Anqiang Wang, Baocai Xing, Ranran Gao, Xianwen Ren,

Li Wang, Xiaojiang Wu, Ji Zhang, et al. Pan-cancer single-cell landscape of tumor-infiltrating t cells. *Science*, 374(6574):abe6474, 2021.


A
Cell type total
Cell type proportion
B
CAF
Cancer
Endothelial
Epithelial
Myeloid
PVL
Plasma
T
0
5
10
15
20
25
30
35
40
45
49
B
UMAP 2
UMAP 1
Cell topic
Cell type
C
Tog genes
Cell type/ Cell topic
SCGB2A1/CXCL8/CD1C
IGF2/SEMA3G/SIGLEC1
CXCL13/KLK3/TAT
MGP/KRT16/KRT23
CEL/CSN1S1/CSN2
MS4A8/PHGR1/SCGB1D2
MUC5AC/CXADRP3/SCGB2A1
FCER2/MUCL1/CD22
RP11−54H7.4/PLCXD3/C6orf15
AADAC/KCNMB2/ADH1B
CXCL13/CCL3/CCR8
FBP2/SPRR1B/SERPINB3
PLN/MT1A/CNN1
MAPT−AS1/RP11−473L15.3/BFSP2
AADAC/TIMP4/TCF15
IL21/DSCAM−AS1/TUBB3
LGALS12/LINC01205/RP11−549K20.1
CTSE/NOL4/IRS4
KRT17/SERPINB4/TYMS
RP11−286H15.1/ACKR1/GPIHBP1
CCL4/RP11−553K8.5/CCL4L2
FDCSP/TCL1A/IGLV3−1
DEFB1/SERPINB3/PCSK1
HBB/AL645728.1/BANCR
HBB/HBA2/HBA1
MS4A1/MUC4/DNAJB1
DLK2/ACTG2/PTN
CST9/TRH/CARD18
APOE/CCL4/PLA2G2D
GRP/EPYC/CST2
TMEFF2/LINC00704/FGF5
B/45
B/20
CAF/7
CAF/22
Endothelial/34
PVL/26
Cancer/48
Cancer/27
Cancer/19
Cancer/2
Cancer/24
Cancer/21
Epithelial/46
Epithelial/17
Plasma/14
Plasma/43
Myeloid/0
T/37
T/23
Myeloid/6
Myeloid/1
Cancer/9
T/39
T/30
B/38
Cancer/32
Cancer/31
Cancer/33
Endothelial/28
B/4
Epithelial/12
10
5
0
−5
−10
D
EPCAM
CD3D
CD68
MS4A1
PECAM1
PDGFRB
UMAP 2
UMAP 1
3
2.5
2
1.5
1
0.5
0
E
B
CAF
Cancer
Endothelial
Epithelial
Myeloid
PVL
Plasma
T
Cell topic proportion


A
UMAP 2
UMAP 1
source cell
target cell
B
Cell pair (i, j)
Ligand of cell i
Receptor of cell i
Ligand of cell j
Receptor of cell j
LR interaction database
Gene expressions aggregated by L-R interactions
Ligand Encoder
Receptor Encoder
softmax
Topic
Ligand Decoder
Receptor Decoder
C
Interaction topics
3M
2M
1M
D
Myeloid
Lymphoid
Stroma
Cancer growth
Cancer metastasis
TME regulation
Endothelial
E
Interaction topic
Endothelial
TME regulation
Cancer metastasis
Cancer growth
Stroma
Lymphoid
Myeloid
Source cell type
B
CAF
Cancer
Endothelial
Epithelial
Myeloid
Plasma
PVL
T
Target cell type proportion
F
Cell types
Interaction topic


A
Endothelial (2)
Myeloid (24)
TME regulation (4)
Lymphoid (22)
Cancer cell
Cancer metastasis (7)
Stroma (18)
Cancer growth (10)
B
Endo.
TME reg.
meta-stasis
Cancer growth
stroma
Lymphoid
Myeloid
Interaction topic proportion
Cell-cell interactions involving 25,835 cancer cells
C
Cancer
Interaction topic proportion
B
CAF
Endoth-
Epithelial
Myeloid
Plasma
PVL
T
Cancer
D
ER+
HER2+
TNBC
E
Receptor
Ligand
A1BG
GFRA1
AQP5
IL20RA
AR
TSPAN15
SDC1
GPNMB
AGR2
VTCN1
ATP6AP2
TSPAN1
ERBB2
KIDINS220
CXCR4
IL10RB
EDN1
ANGPTL4
PIGF
BSG
GSTO1
DEFB1
NMB.2
PIP
TFF1
NTN4.1
NMB.1
GDF15
NTN4
DKK1
LYPD3
NMB
HSP90B1
NAMPT
HSP90AA1
COL6A1
VCAN
FN1
COL6A2
COL6A3
COL3A1
COL1A1
COL1A2
HLA.B
ICAM3
CD58
HLA.C
B2M
HLA.E
HSPA4
HSPA1A
S100A9
PLTP
SERPINA1


A
Ligand
Receptor
Myeloid
Lymphoid
Stroma
Cancer-growth
Cancer-metastasis
TME regulation
Endothelial
B
Cancer metastasis
0.37
0.69
Cancer growth
0.47
0.73
Myeloid
0.41
0.68
Lymphoid
0.30
0.48
Stroma
0.64
0.92
TME regulation
0.64
0.93
Endothelial
0.54
0.80