SPRUCE - Single-cell Pairwise Relationships Untangled by Composite ETM

Sishir Subedi and Yongjin Park

11:43:21 AM, Aug 31, 2022

Introduction

The advancement in single-cell RNA-sequencing (scRNA-seq) has emerged as a new frontier in transcriptomics. The ability to quantify gene expression levels at a single-cell resolution provides a framework to uncover novel cell types and the dynamics of cellular interactions. Studies in the past have shown that the complex cell-cell communication(CCC) among heterogeneous cell populations in the tumour microenvironment (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 provide potential therapeutic insights into cancer.

The traditional approach to studying CCC involves clustering features in low-dimensional space and inferring interactions between clusters of know cell type 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 do not account for intracluster heterogeneous interaction patterns, which is critical in understanding disease complexity Zhang et al. [2021], Tan and Naylor [2022]. Also, 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 CCC relationships as hypergraphs. These methods rely on a priori knowledge of cell type and aggregating cells to calculate communication scores based on the mean expression of 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. Thus, there is a need for alternative approaches that can combine datasets from multiple contexts and untangle CCC at cell-pair resolution. The method also needs to be scalable to large scale datasets and computationally interpretable to gain biological insights.

Embedded topic model (ETM) is a generative deep learning method that uses variational autoencoder architecture to represent data in low-dimension topics with an interpretable topic-to-feature relationship. It has been successfully implemented in natural language processing to extract meaningful topics representing large-scale documents Dieng et al. [2020]. In a recent study, scETM showed that ETM-based techniques efficiently capture the essential biological signals from sparse and heterogeneous single-cell data Zhao et al. [2021]. In this paper, we introduce SPRUCE, Single-cell Pairwise Relationship Untangled by Composite ETM.

Results

Overview of the SPRUCE approach for characterizing interaction patterns between cells

The systematic analysis of cellular composition and cell-cell communication in the tumour microenvironment (TME) is crucial in understanding the complex biological mechanisms behind cancer progression Binnewies et al. [2018]. Two layers of information are critical in deciphering the multifaceted nature of TME:(1) identifying the cellular state of each cell type and (2) the pattern of cell-cell interactions. Accordingly, we developed SPRUCE, a probabilistic deep learning method that groups heterogeneous cell populations into distinct cell states and identifies unique interaction patterns represented in the TME. Briefly, the model is composed of two steps - the first step represents each cell in cell state/topic space, and the next step generates cell pairs based on cell topic assignment and represents each pair in interaction topic space. As the final output, the model extracts pair-wise LR-driven interaction patterns in an unsupervised manner. The cell and interaction topics are informative and biologically interpretable low-dimensional representations of cell states and their dynamic communication pattern.

The model was trained using combined breast cancer data sets with ~155K cells from multiple large-scale studies. First, the model represented all the cells into 50 cell topics, with various topics represented as unique cell states of different cell types, including cancer cells. To represent all the components of TME and capture the extensive crosstalk among tumours, resident cells, and recruited immune cells, we constructed a list of ~25 million neighbour cell pairs where neighbours of each cell represented all cell topics. We then transformed LR gene expression data from each cell pair and represented it in lower dimension of 25 interaction topics. Using SPRUCE, we identified seven different interaction topics dictated by a unique set of LR genes representing tumour-immune, tumour-stromal, and tumour-tumour interactions. The model extracted interaction patterns to represent dynamic cell-cell communication among different cell types present in various cell states in the TME.

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 average 3118 cells (with standard deviation ± 5987). 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 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, which shows distinct clusters for each topic where the majority of cells belong to one of the major types of cells in the dataset. This was further confirmed using cell type lineage canonical markers. 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. We also tried different number of cell topics from 10, 25, 50 and decided to use the 50-topic model because major cell types, especially cancer cells, showed well separated distinct clusters.

Seven common signatures of 25 million cell-cell pairs

We augmented 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. 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 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. The two topics, 22 and 24, captured immune-related interactions. Here, topic 22 was labelled as lymphoid associated topic because top receptors included 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 (ref). Similarly, topic 24 was labelled as myeloid associated 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 (ref).

Topics 10 and 7 represented many oncogenes mutated in cancer- 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. Similarly, topic 7 was cancer-metastasis associated 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.

Further, topic 18 was stroma-associated 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 this topic are COL1A1, COL1A2, COL3A1, and MMP13, and the top receptors are ITGA11 and SCARA5. Similarly, endothelial-associated topic 2 is enriched with genes highly expressed in endothelial cells. Here, endothelial associated ligands such as CD34, ANGPT2, and NID2 and receptors such as APLNR and ESAM are enriched. Likewise, topic 4 was TME-regulation associated topic enriched in genes that promote a conducive environment for cancer growth. The top genes in this topic are KISS1R/KISS1, which play 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.

The top LR genes in the major interaction topics revealed 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 neighbour 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 neighbour cells in that topic. For example, in cancer-growth associated, on average, 68% of neighbour cells for all cell types were cancer cells. Similarly, 49% of neighbour cells in stroma-associated topic were CAF/PVL cells, and myeloid and T cells comprised of 49% and 38% of neighbour cells in myeloid-associated and lymphoid-associated topics, respectively. For endothelial-associated topic, dominant neighbour cell type was endothelial cells with 18%. In contrast, for TME-regulation associated topic, both epithelial and plasms cells were dominant cells consisting of 20% and 22%, respectively. The results show the biological function of topics based on top LR genes identified by the model weights was corroborated by the enrichment of neighbour cell type.

Heterogeneity of breast cancer cells stems from topic-specific interaction patterns

Next, we focused on the interaction patterns of 25835 breast cancer cells distributed among 13 cell topics identified by the cell topic model. The cancer cells manifested all the interaction topics determined by the model, and the interaction pattern depended on the cell type of the neighbour cell. The cancer cells are found to express top LR genes in each interaction topic identified by the model. For immune-related interaction topics and stroma-associated topics, cancer cells show higher expression of ligands than receptors compared to their neighbour cells. But for other interaction topics, especially in cancer-growth associated and cancer-metastasis associated topics, cancer cells and their neighbour cells show similar expression patterns of top LR genes.

We investigated the heterogeneity of different cell types based on their cell topics identified by the cell topic model and their interaction patterns uncovered by the interaction topic model. Here, cancer cells along with epithelial, plasma, and B cells showed heterogeneous interaction patterns compared to myeloid, T, endothelial, CAF, and PVL cell types. For cancer cells, the majority of interactions belonged to cancer-growth associated and cancer-metastasis associated topics where many of the top genes were oncogenes. Here, 55% of the total interactions was cancer-growth associated, and 17% belonged to cancer-metastasis associated topic, while the other five remaining topics consisted of 3-8% of interactions. Among the non-cancer cell types, the dominant interaction topic for myeloid cells was myeloid-associated interaction topic and for T cells it was lymphoid-associated interaction topic. Here, 88% of myeloid cell and 65% of T cell interactions were found to be in respective interaction topic. Similarly, 83% of cell interactions with endothelial cells belonged to endothelial-associated topic, and 77% and 53% of interactions with CAF and PVL cells, respectively were stroma-associated topic. In contrast, plasma, B, and epithelial cells showed a higher mixture non-immune associated interaction topics.

Generally, cell types are defined by the expression of distinct cell markers; however, the transient state of gene expression information may be inadequate in understanding the complex biological interactions of cells with its surrounding in the TME. Therefore, we sought to further explain the heterogeneity of cells based on functional interactions within cell types. Our cell topic model identified 13 cell topics for cancer cells among which many of them showed a distinct pattern of interactions with their neighbouring cells. The cancer-growth associated 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 belonged to cancer-growth associated interaction topic. There were two cell topics in which major interactions were non-cancer like topics- 66% of interactions in cell topic 9 consisted of stroma-associated topic and 60% of interactions in cell topic belonged to endothelial-associated interaction topic. Overall, all the cell topics belonging to T, myeloid, endothelial, CAF, and PVL cell types did not deviate from cell type specific interaction pattern. In contrast, cell topics from plasma, B, epithelial, and cancer cells showed significant variability of interaction patterns among cell topics.

Knowing interaction patterns, cancer types can be refined

We further refined different breast cancer subtypes based on cell topic-specific interaction patterns. We found out that all three different subtypes of cancer cells showed diverse interaction patterns where TNBC cancer cells were more heterogeneous compared to HER2+ and ER+ subtypes. Here, more than 85% of interactions of HER2+ subtype consisted of cancer-associated topics - 82% for cancer-growth and 5% for cancer-metastasis associated topics. Similarly, for ER+ subtype, more than 80% of interactions were cancer-associated topics - 61% for cancer-growth and 20% for cancer-metastasis associated topics. In contrast, TNBC subtype cells were more diverse in interactions, with 42% for cancer-growth associated, 15% for cancer-metastasis associated, 15% for stroma-associated, and 10% for myeloid-associated topics. A closer look at the interaction patterns of different cell topics of cancer subtype found cell topic-specific interaction patterns. Comparatively, the cell topics from HER2+ were more homogeneous than ER+ and TNBC. Our approach was able to identify a specific group of cells (cell topic) within each subtype that showed a distinct pattern of interaction with its neighbouring cells- some groups were highly heterogeneous such as cell topic 21 for TNBC cells, while others were highly homogeneous such as cell topic 2 for HER2+ cells.

Different interaction topics induce subtype-specific gene-gene networks

We generated a topic-specific gene correlation network with significantly expressed LR genes to investigate intercellular communication between cancer cells and their neighbours in different interaction topics. In lymphoid-associated topic, major components of T cell receptor (TCR) complex (CD3D,CD3G, CD2, and CD247) and genes involved in regulating TCR signaling pathway PTPRC, CD45, and CD53 are abundantly enriched, suggesting the regulatory interactions between cancer and T cells. Other genes enriched in this topic are involved in crosstalk between T cells and cancer cells and promote cancer growth and proliferation in the tumour microenvironment. For example, the chemokine receptors CXCR3 and CXCR4 are known to mediate metastasis of breast cancer cells and killer-cell lectin like receptors KLRC1, KLRD1, and KLRF1 are known to restrict T-cell's antitumour immunity. Similarly, immune-modulatory receptors primarily expressed in myeloid lineage cells 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.

The gene interactions in two cancer associated topics showed that genes enriched in cancer-growth associated topic captured the interactions between cancer cells and other cell types that promote its growth, while genes in cancer-metastasis associated topic were active in cancer cell transformation and metastasis. The dominant gene network in cancer-growth associated topic consisted of highly upregulated genes that induces signaling cascades involved in oncogenesis such as PTPRF, FGFR1, ERBB2, and TNFRSF1A. Also, genes known to play an essential role in cellular developmental processes and hijacked by cancer cells, such as LAMP1, ITGB1, RPSA, CANX, ATP6AP2, and MCFD2 are enriched in this topic. Similarly, gene networks in cancer-metastasis topic 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 (EMT).

The stroma-associated topic 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. The genes MCAM, ENPEP, EDNRA, and DCBLD2 that promote blood vessel formation and enhance tumorigenesis are enriched in this topic. Similarly, genes enriched in endothelial-associated topic captures interactions of cancer cells in developing tumour vascular networks, especially in conjunction with endothelial cells. PECAM1, CALCR, ADGRL4, and CD93 genes that are predominantly expressed in endothelial cells and regulate angiogenesis in tumour cells are present in this topic. Additionally, TME-regulation associated topic primarily consisted of genes mixture of enthothelial-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.

References

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.

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.

Mikhail Binnewies, Edward W Roberts, Kelly Kersten, Vincent Chan, Douglas F Fearon, Miriam Merad, Lisa M Coussens, Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C Hedrick, et al. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine*, 24(5):541–550, 2018.

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.

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.

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.

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

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

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.

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, 2021.

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.