RESEARCH

Learning directed acyclic graphs for ligands and receptors based on spatially resolved transcriptomic analysis of ovarian cancer

Shrabanti Chowdhury[1†], Sammy Ferri-Borgogno[2†], Anna P Calinawan[1†], Peng Yang[3,5], Wenyi Wang[3], Jie Peng[4], Samuel C Mok[26] and Pei Wang[1*6]

*Correspondence: pei.wang@mssm.edu
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
Full list of author information is available at the end of the article
[†]Equal contributor[6]co-correspondence

Abstract

To unravel the mechanism of immune activation and suppression within tumors, a critical step is to identify transcriptional signals governing cell-cell communication between tumor and immune/stromal cells in the tumor microenvironment. Central to this communication are interactions between secreted ligands and cell-surface receptors, creating a highly connected signaling network among cells. Recent advancement in in situ-omics profiling, particularly spatial transcriptomic (ST) technology, provide unique opportunities to directly characterize ligand-receptor signaling networks that powers cell-cell communication. In this paper, we propose a novel statistical method, `LRnetST`, to characterize the ligand-receptor interaction networks between adjacent tumor and stroma cells based on ST data. `LRnetST` utilizes a directed acyclic graph (DAG) model with a novel treatment to handle the zero-inflated distribution observed in the ST data. It also leverages existing ligand-receptor regulation databases as prior information, and employs a bootstrap aggregation strategy to achieve robust network estimation. Application of `LRnetST` to ST data of high-grade serous ovarian tumor samples revealed both common and distinct ligand-receptor regulations across different tumors. Some of these interactions were validated through a MERFISH data set of independent ovarian tumor samples. These results cast light on biological processes relating to the communication between tumor and immune/stromal cells in ovarian tumors. An open-source R package of `LRnetST` is available on GitHub at https://github.com/jie108/LRnetST.

Keywords: Spatial transcriptomics data; ligand-receptor network; Hill climbing; Bootstrap aggregation; Prior domain knowledge

1 Background

High grade serous ovarian cancer (HGSC) is the most lethal gynecologic malignancy, with its daunting overall survival rate showing limited improvement over decades [1, 2, 3, 4]. A major obstacle in fully understanding the mechanisms of tumor progression and chemo-resistance in HGSC is its high intra-tumor heterogeneity, comprising both tumor clonal heterogeneity and tissue architecture heterogeneity [5]. The latter is reflected by the heterogeneous stromal and immune cell population in the ovarian tumor micro-environment (TME) [6, 5]. The recent advances in in situ omics analysis have suggested an important link between cell-cell interactions among tumor/immune/stromal cells in TME and tumor progression as well as ther-

apeutic resistance [7]. However, the molecular mechanisms that shape these cell-cell interactions in HGSC are still largely unexplored.

A predominant form of cell-cell signaling is powered through interactions between ligands from one cell and cognate receptors on neighboring cells. Considerable effort has been dedicated to develop tools for exploring these interactions using single-cell RNA-seq (scRNA-seq) data, including CellphoneDB [8], Single-CellSignalR [9], Cellchat [10], ICELLNET [11], CrosstalkeR [12], Nichenet [13], scMLnet [14], and CytoTalk [15]. Despite these endeavors to characterizing intercellular communications with scRNAseq data, the absence of spatial information in scRNAseq data significantly hampers the precision in dissecting ligand-receptor interaction network, given that these interactions occur locally between neighboring cells within the tissue.

The latest development of spatial transcriptomic (ST) profiling technology enables one to map mRNA molecules to a specific location (spot or grid) of a tissue slide in a high-throughput manner [16, 17, 18]. These platforms thus provide an unprecedented opportunity to comprehensively characterize the ligand-receptor interactions among neighboring cells (e.g. those from the adjacent grids on ST slides), which is not feasible based on either bulk or single-cell RNA profiles. In this paper, we aim to characterize the ligand-receptor regulatory network in HGSC using spatial transcriptomic data.

Pioneering effort have been made for inferring cell-cell communication networks while considering spatial information, including `Giotto` [19] and `stLearn` [20]. `Giotto`, a comprehensive open-source toolbox, is designed for analyzing and visualizing spatial transcriptomic data. Within `Giotto`, the Spatially Informed Cell-to-Cell Communication (*spatCellCellcom*) module calculates cell-cell communication scores for each ligand-receptor (L-R) pair between proximal cell types according to the spatial network. Permutation-based p-values and multiple hypotheses adjustment are subsequently computed for each ligand-receptor pair. On the other hand, `stLearn` focuses on the proportion of neighboring spots with upregulated expressions for both the ligand and receptor genes of a given L-R pair. The significant L-R pairs are then obtained by integrating the signals across all spots. Although, both methods incorporate spatial information in the inference of cell-cell communication, neither directly addresses the issue of zero-inflation that is pervasive in ST data. Furthermore, assessing co-expression based on either thresh-holding or marginal correlation is susceptible to the considerable variability present in ST data, leading to a lack of reproducibility, as demonstrated in numerical examples (see Results section).

To address these challenges, we developed `LRnetST` - Ligand-Receptor Network learning based on Spatial Transcriptomics data, a novel tool to construct ligand-receptor networks between adjacent cells of different types based on either multi-cellular or single-cell ST data. The `LRnetST` pipeline begins by introducing the Neighbor Integrated Matrix (NIM), which integrates the spatial information and the molecular information within the ST data. Subsequently, `LRnetST` utilizes a binary variable along-side a continuous variable for every ligand/receptor in the node space. This coding strategy not only addresses zero inflation in ST data, but also enhances the power to detect interactions predominantly signaled through active/inactive statuses of ligands/receptors. To account for variation in library sizes

across grids/cells in one ST experiment, `LRnetST` employs an aggregation framework that combines a bootstrap (bagging) procedure of DAG learning with downsampling based normalization. This aggregation framework provides better control over false edge detection. Moreover, the DAG construction step in `LRnetST` is adaptable to incorporate prior information such as data from existing ligand-receptor databases.

We applied LRnetST to both multi-cellular and single-cellular ST datasets of ovarian cancer. In the multi-cellular dataset, our focus was on detecting ligand-receptor interactions between neighboring spot-pairs enriched with tumor and stroma cells, respectively. In the single-cellular ST data, we examined interactions between tumor cells and other immune or stroma cell types. Specifically, we first applied the proposed LRnetST pipeline to a 10X Genomics ST data of 4 HGSC samples. To guide the construction of DAGs, we incorporate known ligand-receptor regulation information from relevant databases [22] as prior information to constrain the network space. The LRnetST analysis revealed a substantial number of shared ligand-receptor interactions between tumor and stroma cells across four different HGSC samples, outperforming alternative methods. Notably, both LRP1 and ITGB1 were identified as hub nodes, showing connections to multiple ligands in adjacent regions within each of the four patient-specific LR networks. Further applying LRnetST to an independent single-cell MERFISH ST data of ovarian tumors, we were able to confirm several LR interactions, including those between LRP1 in tumor cells and the protease SERPINA1 in the neighboring stromal cells, as well as interactions involving ITGB1 in tumor cells with VEGFA and VCAM1 in neighboring stromal cells. These findings offer fresh insights into the roles of these tumor-relevant genes/proteins in facilitating cell-cell communication within HGSC.

2 Results

2.1 `LRnetST`

To effectively characterize the complicated ligand-receptor regulatory relationship in TME, we proposed to use directed acyclic graphs (DAGs), which has emerged as a valuable tool for inferring gene-gene interactions [21, 22, 23, 24, 25, 26, 27]. However, multiple challenges impede the application of existing DAG models to ST data for constructing ligand-receptor networks. Firstly, integrating both spatial information and -omics profiles concurrently into ligand-receptor network learning is not readily addressed by current DAG methods/software. Secondly, in the sequencing output of an ST experiment, a substantial portion of transcripts is often missing due to a combination of both biological factors (e.g., absence of gene expression in the cell population of that grid) and technical limitations (e.g. limited read depth). The resulting zero-inflated distributions of Unique Molecular Identifier (UMI) counts impose statistical and computational challenges in network inference and mechanistic interpretation. Thirdly, UMI count levels of different spots on one tissue slice are subject to both technical variations such as different library sizes, and biological variations such as cell type heterogeneity. Consequently, proper normalization across spots is essential before undertaking downstream analysis.

To address these challenges, we developed a new DAG model coupled with customized fitting pipeline, called `LRnetST`, for constructing ligand- receptor networks

based on ST data. Here, we provide a concise overview of the key steps in the `LRnetST` pipeline, with comprehensive details of the statistical models available in the Methods section and the complete algorithm pseudocode in Section A.1 of the Additional file 1.

As illustrated in Figure 1A, the `LRnetST` pipeline comprises of four major steps: (1) normalization through down-sampling, (2) construction of the Neighbor Integrated Matrices (NIM) with an additional round of sample bootstrap, (3) estimation of DAGs based on bootstrapped samples, and (4) aggregation of the DAGs into a final network.

Within the `LRnetST` pipeline, the second step contains three key components (Figure 1B). Firstly, for a given normalized ST data matrix (gene by spot/cell) obtained through down-sampling, `LRnetST` integrates the spatial and molecular information through introducing the Initial Neighbor Integrated Matrices (Initial-NIM). This involves combining gene expression information from two neighboring spots/cells into one vector (column). Consequently, evaluating dependence between genes in neighboring spots/cells is equivalent to assessing the dependence between features from the top and bottom halves of the Initial-NIM matrices. Secondly, to account for the high dropout rates in the ST data, `LRnetST` encodes each gene expression using two variables: one binary indicator denoting the expressed or not-expressed status; and a continuous number representing the gene expression level when the UMI count is greater than 0. The resulting data matrix is referred to as NIM, whose feature space is doubled compared to that of the Initial-NIM. This strategy improves the power of detecting dependencies among genes, particularly in the presence of zero-inflated distributions, as demonstrated in the numerical examples below. The third component involves perturbing the sample space of the NIM through bootstrap, where the resulting matrix serves as the input of the subsequent DAG construction. Notably, the ensemble and aggregation framework in `LRnetST` through integrating down-sampling-based normalization, NIM construction, and the bootstrap-based network inference, not only effectively accounts for the varying UMI count levels across different ST spots/cells, but also enhances the robustness of the DAG estimation.

The third step of `LRnetST` contains a novel DAG model customized to the NIM matrices. Specifically, the DAGs are constructed by minimizing a score function, amounting to maximize a BIC penalized joint likelihood of all the random variables in the graph. Similar to a previous work [28], continuous variables conditioned on their parent nodes are modeled with Gaussian distributions, while a logistic model is employed to characterize the dependency of binary variables on their parent nodes. In addition, from the definition of the binary and continuous variable of each gene, it's obvious that the binary variable is a parent node of its corresponding continuous variable and the likelihoods of the continuous variables are evaluated only for spots where the corresponding binary nodes take the value 1 (i.e., when the continuous variables are defined). Moreover, when building the DAG based on the NIM matrix, since edges of interests are between genes from different spots, only edges between features from the top (i.e., index spots) and the bottom (i.e., the paired neighbor spots) of the NIM matrix are considered. Moreover, `LRnetST` incorporates existing ligand-receptor databases as prior information to further constrain the search space of directed edges during DAG fitting by the hill climbing algorithm.

In the last step, the derivation of the final aggregated DAG is achieved by searching for a DAG that minimizes an average structural Hamming distance to the DAGs in the ensemble built from the previous steps (see Methods).

Please note that LRnetST is a versatile tool and can be applied to both multi-cellular and single-cellular ST datasets. In the real application detailed in this manuscript, we demonstrated LRnetST's utility by analyzing both a multi-cellular and a single-cell ST datasets of ovarian cancer. For the multi-cellular data, our primary objective was to identify ligand-receptor interactions between neighboring spot-pairs enriched with tumor and stroma cells respectively. And to validate the findings, we extended the analysis to an independent single-cell ST dataset of ovarian cancer, where we investigated LR interactions between tumor cells and other immune or stroma cell types. For more details, please refer to the following sections.

2.2 Evaluation of DAG construction based on synthesis data sets

We utilized synthesis data sets (Initial-NIM and NIM matrices) to evaluate the performance of the DAG construction step in `LRnetST`. When generating the synthesis data sets, we considered various scenarios with different numbers of genes, (true) edges, and Index-Neighbor spot pairs (i.e., samples sizes). See the Method for details. Specifically, we considered two different sample sizes n=200 and 400; and varied the number of genes in true DAG from 100 – 600 with different number of true edges, thus generating 12 DAG topology (Figure S1A-F). For each DAG and a given sample size, we generated 100 independent data sets. The performance evaluation was conducted by assessing the powers or true positive rates (TPRs), false discovery rates (FDRs) and the F1 scores for two scenarios: (i) detection of skeleton edges (i.e., without direction); and (ii) edge detection with direction.

We compared the performance of DAG construction based on initial-NIM and NIM to assess the advantage of using the paired binary and continuous nodes to facilitate the modeling of zero-inflation in ST data. Moreover, we compared the performance of `LRnetST` to alternative DAG construction methods including `DAGBagM` [28], `DAGBagM_C` [28], and `bnlearn` [29]. Note, `DAGBagM`, which can handle binary and continuous variables simultaneously, was applied to NIMs; while `DAGBagM_C`, which model continuous variables, were applied to the initial NIM before introducing the binary nodes. `Bnlearn` was also applied to the initial NIM after converting it to binary variables using median cut-off. Note that, we performed bootstrap for all methods except `bnlearn` due to its prohibitive computational cost.

The results in Table A.1 in Additional file 1 and Figure S1G-H indicate superior performance of `LRnetST` compared to the other methods across all considered scenarios, illustrating the advantages of including the binary nodes in NIM in the `LRnetST` pipeline. Furthermore, the application of both `LRnetST` and `DAGBagM` on NIM underscores the superior performance of `LRnetST` over `DAGBagM`, emphasizing the advantages derived from the customized treatment of likelihood terms associated with paired binary and continuous nodes in `LRnetST`. In the end, as expected, for all four methods, we observed improved performances as the sample sizes (n) increased; while declined performance with the increase of the number of genes (p).

2.3 Revealing cell-cell interaction in ovarian cancer with LRnetST

2.3.1 Applying LRnetST to the 10X Genomics ovarian cancer ST data

For each tumor sample in the 10X Genomic ovarian cancer ST data set (see the Data Description under Methods Section), we applied LRnetST to construct tumor-specific ligand-receptor networks. Firstly, to identify tumor or immune/stromal cell-enriched spots, we estimated the cell-type proportions in each grid (spot) by performing deconvolution analysis using *runDWLSDeconv* implemented in Giotto with the scRNAesq data of ovarian tumors from [30] as the reference (see Methods). Based on the cell-type proportion estimates, we then inspected the tissue and cell morphology in high resolution Hematoxylin and eosin stain (H&E) images, and manually annotated the ST grids (spots) enriched with tumor or stromal cells (Figure S2A-D, Methods).

To construct the NIM, we screened for tumor cell-enriched (index) spots with neighboring (within 2-grid-units on the ST slices) stromal cell-enriched spots. Using these index-tumor and neighboring-stroma spot-pairs, we assembled the Initial NIM and then NIM as described in the previous section (see Figure 1). Note, in this process, we focused on the genes documented in the ligand-receptor interaction database [31], and further filtered out those detected in less than 50% of the spots in one sample. The median of the total UMIs per spot in each sample was used for the down sampling normalization in LRnetST. Table A.2 in Additional file 1 gives the detailed numbers of the tumor and stromal spots, as well as the numbers of adjacent tumor-stromal-spot pairs in each sample. The table also lists the numbers of ligand/receptor genes with adequate UMI data points to construct the network in each sample.

Using LRnetST, we built four ligand-receptor networks and obtained 422, 170, 334, and 210 ligand-receptor regulation edges for samples A4, A5, A10 and A12, respectively (Figures S4A-B, S5A-B, Table S1). Both common and distinct interaction patterns were observed across the four tumor samples.

Comparing LRnetST with stLearn and spatCellCellcom

To compare the performance of LRnetST with other ligand-receptor network learning methods, we also applied stLearn and spatCellCellcom on the same data set. For pre-processing including normalization, we followed the respective pipelines of these two methods implemented in the corresponding R (spatCellCellcom implemented under Giotto) and python (stLearn) packages. Similar to LRnetST, we only considered the genes documented in the ligand-receptor database and survived the missing filtering cut-off of 50%. Both stLearn and spatCellCellcom provided adjusted p-values for all possible interactions between ligands and receptors. By thresh-holding the adjusted p-values, stLearn inferred 113, 111, 103 and 36 ligand-receptor regulation edges (Table S1), respectively; while spatCellCellcom identified 188, 123, 330 and 66 edges (Table S1), respectively, for samples A4, A5, A10 and A12.

Reproducibility of the edges across 4 tumor samples

We first evaluated how "reproducible" the inferred ligand-receptor (LR) edges are across different tumor samples. Intuitively, methods detecting more shared (reproducible) edges across tumors shall enjoy better power for detecting meaningful biological interactions. As summarized in Figure 2A-C (vertical bar heights indicating number of common edges shared across at least 2 out of 4 tumor samples), `LRnetST` inferred many more shared edges across multiple samples than the other two methods. Specifically, `LRnetST` inferred 185 common edges, while `spatCellCellcom` inferred 67 and `stLearn` inferred 48 common edges across the tumors. Note that, for counting the common edges, we considered undirected networks and further collapsed the binary and continuous component of the genes. For example, `LRnetST` identified 4 edges shared by all four tumors, whereas the other two methods failed to identify any such edges. Furthermore, we assessed whether the overlap between LR edges for a pair of tumors significantly surpassed what would be expected by random chance, as outlined in the Methods section. The resulting adjusted p-values are presented in Figure 2D. The outcomes from `LRnetST` exhibited significant overlap for all six pairs of tumors, whereas the results from `spatCellCellcom` and `stLearn` demonstrated significant overlap for only half of the pairs. This suggests the higher level of reproducibility of `LRnetST` compared to the alternatives.

We then investigated the hub nodes in the LR networks by the three methods (Figure S3). Hub nodes are those with higher connections in a network, and thus are often more likely to play an essential role in the regulatory network [32]. Labeling the top 30 nodes with the largest degrees in each network as the hub nodes, we identified three common hub nodes, LRP1, ITGB1 and ITGAV, across all four tumors based on `LRnetST`'s results. On the contrary, only ITGB1 was detected as a common hub node by `spatCellCellcom` and `stLearn`.

Interestingly, in `LRnetST` results (Figure S3A), both LRP1 levels in tumor cell-enriched spots (LRP1_tumor) and that in the stroma cell-enriched spots (LRP1_stroma) appeared to be hub nodes in all LR networks. The degrees of the LRP1 tumor nodes are 8, 18, 7, and 4 in the four samples, respectively, while the degrees of the LRP1_stroma nodes are 13, 22, 5, and 4, respectively. Of all the edges associated with LRP1, the interaction between LRP1_stroma and MMP9_tumor (MMP9 levels in tumor cell-enriched spots) is the only common edge identified in all four tumor samples (Figure 2H and Figures S6B, D, F). Previous studies have highlighted the active involvement of the LRP1-MMP9 interaction in processes such as the migration of epithelial ovarian cancer (EOC) [33] and other cancer types, as well as in the development of cerebral edema (see Discussion).

Moreover, a few interactions involving LRP1 were exclusively detected in tumors from chemo-sensitive or chemo-refractory patients (Figure 2G-H, and Figure S6A-F). For example, in the two chemo-sensitive patients, LRP1_Tumor is inferred to interact with C1QB from the nearby stroma spots, which supports the potential role of LRP1-C1QB crosstalk network in modulating immune response in HGSC (see Discussion). On the other hand, LRP1 Tumor is detected to interact with SERPINA1, a serine protease inhibitor, from the nearby stroma spots only in samples from the two chemo-refractory patients. This regulation implies a potential connection between LRP1 and the protease [34].

Apart from LRP1, ITGB1 is another hub node inferred to interact with multiple genes in all four tumors (Figure 2E-F, and Figure S7A-D). Specifically, for both

chemo-refractory patients, ITGB1_tumor was detected to interact with VCAM1 and VEGFA from the nearby stroma spots. At the same time, ITGB1_stroma was connected with multiple genes, including LAMC2 and VEGFA, from the nearby tumor spots. Multiple literatures reported that the intercellular communication happens through the interaction between ITGB1 and VEGFA [35]. In addition, ITGB1 and VCAM1 are both cell-adhesion molecules that influence immune response and anti-immunity [36] (see Discussion).

To further assess the performance of various methods, we then applied `LRnetST` to a cell-level MERFISH data to validate some of the detected LR interactions.

2.3.2 `LRnetST` applied on MERFISH data for validation

We studied four MERFISH data sets [37] derived from four tumor slices of two ovarian cancer patients (one slice from one patient and three slices from another patient). MERFISH, a spatially resolved single-cell transcriptome profiling technology, offers cell-level resolution, enabling detailed monitoring of LR interactions between individual cells. This granularity is particularly advantageous for overcoming the challenge of cell type mixtures within spots encountered in 10X genomics ST data. However, owing to sequencing limitations inherent to MERFISH, only a limited set of transcripts, in the range of few hundreds, can be detected and quantified. This restriction significantly constrains the ability to systematic characterize LR interactions, underscoring our use of MERFISH data for validation rather than exploration purposes.

For each tumor slice, after pre-processing, clustering and cell type labelling (see Method), we identified neighboring tumor-stroma or tumor-immune cell pairs and constructed the NIM for applying `LRnetST`. The number of adjacent tumor-stromal/immune cell pairs for each tumor slice are given in Table A.3 in the Additional file 1. Figure 3A illustrates the spatial distribution of tumor cells and macrophage cells in one tumor slice. While there are thousands of epithelial (pink) and macrophage (light blue) cells in the tumor slice, we focused on the neighboring tumor-macrophage pairs that are in proximity with each other (red epithelial cells and dark blue macrophage cells) for the `LRnetST` analysis. Similar plots illustrating spatial distribution of tumor cells and other stroma cell types, including T-cells, fibroblast and endothelial, are shown in Figure S8A-C. With `LRnetST`, we built a collection of ligand-receptor networks one for each cell type pair (i.e., tumor-macrophage, tumor-T-cell, tumor-fibroblast, and tumor-endothelial), and each tumor slice (Table S2). Figure 3B-E illustrates subsets of the inferred networks considering different neighboring cell type pairs for one tumor slice.

We then assessed the validation of the inferred LR interactions from the 10X data using `LRnetST` in the MERFISH datasets. Specifically, we focused on the 185 reproducible edges, identified in at least two out of the four 10X data sets by `LRnetST`. Among these, only 27 edges had both nodes measured in the MERFISH datasets. Out of these, 19 edges were confirmed in the LR network constructed based on the MERFISH data, i.e., the corresponding edge was detected in at least one of the LR networks derived from MERFISH data. The validated edges in `LRnetST`'s result are illustrated in Figure 4A. Specifically, interaction between LRP1_tumor and SERPINA1_stroma was confirmed in multiple MERFISH data sets (Figure 3B-C).

Moreover, the bi-directional interaction between ITGB1 tumor — LAMC2 stroma and ITGB1 stroma — VEGFA tumor was also confirmed across multiple tumor slides (Figure 3B-E), further strengthening the confidence of these LR interactions in ovarian tumor tissues.

In parallel, we performed the same analysis using `spatCellCellcom` and `stLearn`. As illustrated in Figure 4B-C, only 3 and 1 edges in the results of `stLearn` and `SpatCellCellcom` were validated in the MERFISH data, respectively. Note that the interaction between ITGB1 and VEGFA was inferred by all three methods (Figure 4B-C), suggesting the robust (strong) signal of this LR interaction in ovarian tumor tissues.

Discussion and conclusion

To gain insights on the molecular basis of tumor micro-environment, in this paper, we introduce a new method — `LRnetST` to infer ligand-receptor interaction networks among adjacent cells using spatial transcriptomic data. `LRnetST` employs novel strategies to address the spatial structure and high dropout rates that are uniquely present in the ST data. We demonstrated the benefit of the strategies of `LRnetST` for modeling the zero-inflated ST data over alternative approaches on simulation examples. We then applied `LRnetST` to 10X ST datasets of four HGSC cancer samples and constructed patient specific ligand-receptor networks. The results reveal both common and distinct ligand-receptor regulations across different patients. A subset of these interactions were validated through a MERFISH data set of independent ovarian tumor samples.

We also compared performance of `LRnetST` and alternative tools on the ovarian ST datasets. The results highlight `LRnetST`'s superior capability in detecting reproducible signals that hold greater biological significance.

In the results of `LRnetST`, LRP1 is identified as a hub node in all four ligand-receptor networks in the 10X data results. LRP1 is reported to make critical contribution to many processes that drive tumorigenesis and tumor progression [38, 39], and has been proposed to be an important diagnostic and prognostic biomarker of epithelial ovarian cancer [33]. Besides the previously reported role of LRP1 in ERK signaling pathway through interaction with MMP9 [33], which was also supported in our results (Figure 2G), we further revealed strong interaction between LRP1 in tumor cells and SERPINA1 (AAT) as well as C1QB in stroma cells (Figure 3B-C, 4A, and S6A, S6C, S6E). The interaction between LRP1 in tumor cells and SERPINA1 in stroma cells was detected in the two short term survivor samples (PFS $<$ 6 months) in the 10X data sets, and was consistently confirmed across multiple tumor samples in the MERFISH data (Figure 3B-C, 4A). While the role of LRP1 and SERPINA1 interaction in the context of cancer remains largely unexplored, existing research in cardiovascular studies has highlighted its significance in an anti-inflammatory mechanism [40]. By analogy, one may hypothesize that the interaction between LRP1 in tumor cells and SERPINA1 in stroma cells could contribute to an immunosuppressive effect within the tumor microenvironment.

Interaction between LRP1 and C1QB, on the other hand, was only detected in the two long term survivor samples in 10X data sets. C1QB is a pattern recognition molecule of innate immunity and can be locally synthesized by macrophages and

dendritic cells [34]. The connection between LRP1 and C1QB detected in the two chemo-sensitive patients suggests a potential role of LRP1-C1QB crosstalk network in modulating immune response which may then contribute to the chemo-treatment effectiveness in HGSC patients. Unfortunately, C1QB was not detected in the MERFISH data sets, so validation of this interaction using additional resources is warranted as future studies.

Another hub node detected in all four tumors based on 10X data sets was ITGB1, which has been inferred to interact with VEGFA and VCAM1 in the two chemo-refractory patients. These interactions were also validated on the MERFISH data. Intercellular communication mediated via interaction between VEGFA and ITGB1 has been reported to contribute to the precise coordination of multiple biological processes, including development, differentiation, and inflammation [35]. Recently, interactions through VEGFA and ITGB1 was reported to underlie the crosstalk between epithelial cells and fibroblasts relating to tumor inflammation in Pancreatic ductal cancer [41]. ITGB1 and VCAM1 are both cell-adhesion molecules that influence immune response and anti-immunity. VCAM1 mediates intercellular adhesion by specific binding to the integrin ITGB1 on leukocytes, and thus this interaction plays a pathophysiologic role in leukocyte emigration to sites of inflammation [36]. The interaction between ITGB1 and VCAM1 has also been reported in the context of cell migration between brain endothelial cells (BECs) and central memory T cells (Tcm) [42]. However, our analysis, for the first time, suggests the potential roles of the cross-talk between ITGB1, VEGFA and/or VCAM1 among ovarian cancer patients.

In conclusion, spatial transcriptomic data offers an unprecedented opportunity to unravel the molecular mechanisms underlying cell-cell interactions within the tumor microenvironment. The `LRnetST` method, proposed in this study, proves to be a valuable tool for constructing ligand-receptor interaction networks based on spatial transcriptomic (ST) data. Similar to other high-dimensional network construction methods, the performance of `LRnetST` may be impacted by the dimension of the gene space, as demonstrated in our simulation study. This underscores the significance of leveraging prior domain knowledge to constrain the Directed Acyclic Graph (DAG) search space when applying `LRnetST`, as exemplified in our ovarian cancer ST data analysis through the use of existing ligand-receptor networks. The outcomes of these enhanced analyses promise to unveil crucial contributors driving cell-cell interactions, potentially leading to the identification of new predictive biomarkers or therapeutic targets. An R package of `LRnetST` is made available as a github repository `https://github.com/jie108/LRnetST`.

Methods

Notation and background

We first provide a brief review of directed acyclic graph (DAG) models. A directed acyclic graph $\mathcal{G}(\mathbb{V}, \mathbb{E})$ consists of a node set $\mathbb{V}$ and an edge set $\mathbb{E}$ with directed edges from *parent* nodes to *children* nodes, while there is no cycle in the graph. In the DAG models, the node set $\mathbb{V}$ corresponds to a set of random variables, and the edge set $\mathbb{E}$ encodes the conditional dependence relationships among these random variables. DAG structure learning amounts to identifying the parent set (also referred to as neighborhood) of each node in the graph. Although different DAGs could encode the same set of conditional dependencies (which form an equivalent class of DAGs), it is shown that, two DAGs are equivalent if and only if they have the same set of skeleton edges and v-structures ([43]). The skeleton edges are obtained by removing directions from the directed edges and a v-structure is a triplet of nodes (x_1, x_3, x_2), such that $x_1 \to x_3 \in \mathbb{E}$, $x_2 \to x_3 \in \mathbb{E}$, and x_1, x_2 are not adjacent. There are mainly three classes of methods for DAG structure learning, namely, *score-based* methods (e.g., [44]), *constraint-based* methods , e.g., PC algorithm (`PC-alg`) [43, 45, 46] and *hybrid* methods, e.g., Max-Min Hill Climbing (`MMHC`) [47].

LRnetST

We present a new tool - LRnetST - for learning ligand-receptor interaction networks based on spatially resolved transcriptomic data. First, LRnetST integrates the spatial and the molecular information in the ST data by introducing Neighbor Integrated Matrices (NIM). Second, to account for the high dropout rates in the ST data, LRnetST codes each gene's active/inactive (expressed/not-expressed) status by a binary variable and uses a continuous variable to represent one gene's expression level given the gene is active (expressed). Moreover, to account for the varying UMI count levels across different ST spots, LRnetST incorporates an aggregation framework that conveniently couples the downsampling based normalization with a bootstrap based network inference procedure.

Finally, LRnetST can take into account prior information of interactions, such as those from existing ligand-receptor databases by constraining the DAG search space to the documented edges among the ligands and receptors in the databases. **Figure** 1 illustrates the flow of the LRnetST pipeline. See also **Algorithm** in section A.1 in the Additional file 1. In the following, we elaborate the major steps.

Neighbor Integrated Matrix (NIM)

In the LRnetST pipeline, we first derive an initial Neighbour Integrated Matrix (Initial-NIM) to integrate the spatial and the molecular information in ST data. Since our goal is to characterize whether/how receptors in non-tumor cells in the micro-environment are affected by the ligands from the neighboring tumor cells or vice versa, we choose to focus on pairs of adjacent ST spots enriched of tumor and stromal/immune cells, respectively. Specifically, for each tumor sample, ST spots are first classified into two classes: enriched or not enriched of tumor cells. We then identify the subset of tumor-cell-enriched spots sitting on the boundary of the tumor regions (i.e. being adjacent to spots enriched of non-tumor cells on the ST slice). We refer to these spots as the Index Spots and refer to their closest non-tumor-cell-enriched spots as their Neighbor Spots. Denote the total number of Index Spot and Neighbor Spot pairs as n, and the number of genes under consideration as p. In the next step, we derive an Initial-NIM by stacking the rows of the gene expression matrix of the Index Spots, which has a dimension of $p \times n$, with the gene expression matrix of the corresponding Neighbor Spots, which also has a dimension of $p \times n$. The resulting Initial-NIM has an expanded feature space of $2p$ rows (**Figure** 1). Finally, we convert the *Initial-NIM* into *NIM* by replacing the expression Z of a gene with a pair of features, X and Y to facilitate the modeling of zero-inflation in the ST data (Please see the next subsection on "Accounting for zero inflation in ST data"). The resulting NIM has a dimension $4p \times n$. See **Figure 1B** and the pseudo code of the algorithm outlined in section **A.1** of the Additional file 1 for more details.

Accounting for zero inflation in ST data

Similar to scRNA-seq data, ST data has a high dropout rate, i.e. only a fraction of the transcriptome can be detected in the sequencing result of each ST spot. To facilitate the modeling of such zero-inflated distributions, LRnetST uses two nodes (variables) to represent each gene: a binary node and a continuous node.

First, we use Z_{ij} to denote the downsampled UMI count of the i_{th} gene in the j_{th} ST spot (see *Section 3.2.4* for details on downsampling normalization). Note, $Z_{ij} = 0$ usually corresponds to the situation where either the i_{th} gene is not expressed, or its expression level is low in the j_{th} spot and thus is difficult to be detected in the sequencing experiments. Therefore, we view the event of $Z_{ij} = 0$ as a "less-active" status of the gene and model this status directly. Specifically, the binary node, denoted as Y_{ij}, represents the detection status of a gene:

$Y_{ij} = 1$, if $Z_{ij} > 0$ (i.e., the gene is detected); and $Y_{ij} = 0$, otherwise (i.e., the gene is not detected).

Given the gene is detected. (i.e., $Y_{ij} = 1$), the continuous node, denoted as X_{ij}, represents the normalized gene expression/abundance:

$$X_{ij} = Z_{ij}, \text{ if } Y_{ij} = 1; \quad \text{and } X_{ij} = NA, \text{ if } Y_{ij} = 0.$$

Then, by including both nodes, X_{ij} and Y_{ij}, LRnetST not only nicely accounts for the zero inflation in the data distribution, but also achieves better power in detecting those interactions that are largely signaled through the active/less-active statuses of genes.

Downsampling normalization and bootstrap model aggregation

LRnetST also employs an aggregating framework that couples a bootstrap (bagging) procedure of DAG learning with the downsampling based normalization to achieve better control of false edge detection and to account for the varying library sizes or cell type compositions across different tissue spots.

We apply downsampling normalization on the original gene expression data by fixing the total UMI at the median level of the total UMI counts across all ST spots. We then generate $B = 100$ downsampled gene expression matrices ($[Z]$), and construct $B = 100$ NIMs; also see **Figure 1(b)**) accordingly. We then obtain one bootstrap resample on each NIM through sampling with replacement of the Index/Neighbor Spots pairs (i.e., the columns of the NIM). Finally, we learn one DAG based on each bootstrap resample using the method described in the previous subsection.

The resulting ensemble of DAGs are then aggregated following the aggregation procedure implemented in DAGBagM [28], where the hill climbing algorithm is (again) used to search for the DAG that minimizes an aggregation score based on the *structural Hamming distance (SHD)* while maintaining acyclicity.

Score function and optimization

In LRnetST, we learn DAG by minimizing a score function through a greedy search algorithm (*hill-climbing*), where the score of the continuous nodes are calculated only using data from spots on which their binary nodes take the value 1 (i.e., X is not NA). This makes the binary nodes natural parents of the corresponding continuous nodes of the same gene. Specifically, for a continuous node X, at each search step, its score is calculated by regressing X onto its current parent(s), using only the samples where X is non-zero: For a given

graph $\mathcal{G}$, $score_{BIC_X} := n_X \log(RSS_X^{\mathcal{G}}/n_X) + |pa_X^{\mathcal{G}}| \log(n_X)$, where, $RSS_X^{\mathcal{G}}$ is the residual sum of squares, n_X is the number of samples where X is non-zero and $pa_X^{\mathcal{G}}$ denotes the parent set of X in graph $\mathcal{G}$. To obtain the score for a binary node Y, we regress Y onto its current parent set $pa_Y^{\mathcal{G}}$ through logistic regression:

$$score_{BIC_Y} := -2\left(\sum_{l=1}^{n} I(Y_l = 1)\log(\hat{p}_l) + \sum_{l=1}^{n} I(Y_l = 0)\log(1-\hat{p}_l)\right) + |pa_Y^{\mathcal{G}}|\log(n),$$

where $\hat{p}_l = P\left(Y_l = 1 | pa_{Y,l}^{\mathcal{G}}\right) = \frac{\exp(\hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}^T pa_{Y,l}^{\mathcal{G}})}{1+\exp(\hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}^T pa_{Y,l}^{\mathcal{G}})}$, $\hat{\gamma}_0$ is the fitted intercept and $\hat{\boldsymbol{\gamma}}$ is the vector of the fitted coefficients in logistic regression, and n is the sample size.
Finally, the score of a graph $\mathcal{G}$ is the summation of individual node scores:

$$score_{BIC}(\mathcal{G} : \mathbb{D}) = \sum_{X:\textbf{continuous nodes}} score_{BIC_X} + \sum_{Y:\textbf{binary nodes}} score_{BIC_Y},$$

where $\mathbb{D}$ denotes the data. We then search for the DAG that minimizes the above score function using an efficient implementation of the hill climbing algorithm that is modified from one of our recent works `DAGBagM` [28] – a DAG method for mixed types of nodes.

Simulation Studies

We conduct simulation experiments to examine the performance of `LRnetST` and further compare it with three alternative methods: `DAGBagM`, `DAGBagM_C` [28] and `bnlearn` [29]. Specifically, `DAGBagM_C` and `bnlearn` are applied to the Initial Neighbor Integrated Matrix (**Figure** 1b) before introducing the binary nodes. `DAGbagM`, similar to `LRnetST`, is applied to the NIM after introducing the binary variables in the node space. However, for `LRnetST.DAGbagM`, X is replaced with Z (NA in X is replaced with 0) in NIM, and all samples are used to calculate the scores for the continuous nodes; whereas `LRnetST` uses the converted X so that only samples with $Y = 1$ are used to calculate the scores for the continuous nodes. By comparing the performance of `LRnetST` to these alternative strategies, we can characterize the benefit of using both binary and continuous nodes to facilitate the modeling of zero-inflation in ST data.
We perform simulation experiments for various combinations of the number of genes p, the number of edges ($||E||$), and the number of Index-Neighbor spot pairs (n). Given p genes, we first introduce $4p$ nodes as following: For a given gene i, we use two nodes (Z_i, Y_i) to indicate its expression level and active status in the Index Spot, and $(\tilde{Z}_i, \tilde{Y}_i)$ to indicate its expression level and active status in the Neighbor Spot. We then randomly choose m edges (while excluding cycles) between one node corresponding to Index Spot and another node corresponding to Neighbor Spot. The edges are of the following patterns: (i) $Z_i \to \tilde{Z}_j$; or $\tilde{Z}_i \to Z_j$ or (ii) $Z_i \to \tilde{Y}_k$; or $\tilde{Z}_i \to Y_k$ or (iii) $Y_k \to \tilde{Z}_j$; or $\tilde{Y}_k \to Z_j$ or (iv) $Y_k \to \tilde{Y}_s$; or $\tilde{Y}_k \to Y_s$. This process creates a DAG $\mathcal{G}^*$ with $p^* = 4p$ nodes and m edges. We also introduce directed edges from each binary node $Y/\tilde{Y}$ to its corresponding continuous node $Z/\tilde{Z}$ and pass them as $whitelist$ (i.e., edges always kept in the search steps) while learning the DAGs using `LRnetST` and `LRnetST.naive`.
Given $\mathcal{G}^*$ and sample size n, i.i.d. samples of continuous nodes are generated according to *Gaussian linear mechanisms*: $Z_i = \sum_{pa_{Z_i}^{\mathcal{G}}:continuous} \beta_{ij}\tilde{Z}_j + \sum_{pa_{Z_i}^{\mathcal{G}}:binary} \eta_{ik}\tilde{Y}_k + \eta_i Y_i + \epsilon_i, \quad i = 1, \cdots, p$, where ϵ_i's are independent Gaussian random variables with mean zero and variance σ_i^2. The coefficients β_{ij}s, η_{ij}s, etc., are uniformly generated from $\mathcal{B} = [-0.5, -0.3] \cup [0.3, 0.5]$ and the noise variances σ_i^2s are chosen such that for each node the corresponding *signal-to-noise-ratio (SNR)*, defined as the ratio between the standard deviation of the signal part and that of the noise part, is within $[0.5, 1.5]$. Moreover, i.i.d. samples for the binary nodes are generated by logistic regression models:
$logit(P(Y_k = 1)) = \gamma_0 + \sum_{pa_{Y_k}^{\mathcal{G}}:continuous} \gamma_{kj}\tilde{Z}_j + \sum_{pa_{Y_k}^{\mathcal{G}}:binary} \delta_{ks}\tilde{Y}_s, \quad k = 1, \cdots, p$. Binary nodes without any parent are generated from Bernoulli(0.5). $\tilde{Z}_i$s and $\tilde{Y}_k$s are generated similarly. Finally, for each gene $1 \le i \le p$ and on each sample $1 \le l \le n$, we define $X_{il} = Z_{il}$ if $Y_{il} = 1$; Otherwise $X_{il} = NA$. $\tilde{X}_{il}$s are defined similarly.
For each simulated dataset, we fit `LRnetST` to the $4p \times n$ stacked X, $\tilde{X}$, Y and $\tilde{Y}$ matrix (see **Figure** 1) and obtain the estimated DAG. For the purpose of comparison, `DAGBagM` is applied to the $4p \times n$ stacked Z, $\tilde{Z}$, Y and $\tilde{Y}$ matrix, with the same $whitelist$ used as mentioned above. On the other hand, `DAGBagM` and `bnlearn` are applied to the $2p \times n$ stacked Z and $\tilde{Z}$ matrix with only continuous nodes. `DAGBagM` is a previously developed algorithm by us where the same aggregation procedure has been implemented as in `LRnetST`; whereas we use `bnlearn` without any aggregation (due to its computational slowness). Both `DAGBagM` and `bnlearn` are score-based DAG learning methods using the hill-climbing algorithm.
We consider simulations with different numbers of genes ($p = 100, 200, 300, 400, 500$, and 600) and different sample sizes (i.e. numbers of Index-Neighbor Spot pairs: $n = 200$ and 400). The true DAGs corresponding to different p are shown in **Figure** S1A-F. For each DAG and a given sample size, we generate 100 independent data sets. We then evaluate the performance of `LRnetST` and other methods by assessing their powers or true positive rates (TPRs), false discovery rates (FDRs) and the F1 scores for: (i) detection of skeleton edges (i.e., without direction); and (ii) edge detection with direction. Power (or TPR) is calculated as the ratio between the number of correctly identified edges in the estimated DAG to the number of total edges in the true DAG. FDR is calculated as the ratio between the number of falsely identified edges in the estimated DAG to the number of total edges in the estimated DAG. F1-score is calculated as $F1 = 2 \times$ precision $\times$ recall/(precision+recall), where precision=1-FDR and recall = power. All numbers are averaged over the 100 independent replicates.

Data Description and Processing

10X Genomics ST Data

Spatial transcriptomic analysis was performed on four fresh frozen treatment-naive advanced stage HGSC samples using the 10X Genomics platform as described in our previous work [48]. Among these samples, two (A4 and A5) were derived from chemo-sensitive patients with extended progression free survival time (>10 years); and the remaining two (A10 and A12) originated from chemo-resistant patients with short progression free survival time (<6 months). In the ST experiment, each ST spot contained 20-50 cells, and 930-1007 spots were profiled per tumor slice.

Expression levels of genes in each spot were measured using counts of unique molecular identifier (UMI). Specifically, the output UMI count matrix for each sample (spot) is obtained using ST pipeline (version 1.80.1) following the recommended default parameter setting from FASTQ files [49]. Human genome reference (GRCh38) was used for read alignment and corresponding genome index was generated using STAR2 (version 2.7.9a) [50]. In the preprocessed ST data, the number of genes with non-zero UMI ranged from 68 - 9364 across all grids in all four samples. The median of the total-UMI per grid is 87633.5, 1066, 13318, and 7473 for the four tumor samples, respectively.

For the 10X genomics ST datasets, since each spot is a mixture of several cell types, in order to identify tumor or immune/stromal cell-enriched spots, we estimated the cell-type proportions in each grid (spot) by performing deconvolution analysis using the function *runDWLSDeconv* implemented under `Giotto` with the scRNAesq data of ovarian tumors from from five HGSC patients [30] as the reference. Specifically, we used the gene signatures upregulated in the cell-types Adipose, BCL, Endothelial, Epithelial, Fibroblast, Macrophage, Monocyte, NKC and TC8 from the above-mentioned scRNAseq data and used the *R* function *runDWLSDeconv* implemented under `Giotto` to estimate the cell-type proportions in the 10X ST data. Based on the proportions, we only kept the spots with (1) high Epithelial cell percentage (low immune/stromal cell percentage) and (2) high immune/stromal cell percentage (low Epithelial cell percentage) and filtered out the remaining. Further we inspected the H&E image of the tissues to annotate the tumor enriched spots from (1) and spots enriched with immune/stromal cells from (2).

For each tumor spot, we further identified the adjacent or neighboring immune/stromal spot that are at most 2-grid distance apart from the tumor spot. In this way, we filtered out some of the tumor spots that did not have any immune/stromal spot within 2-grid distance.

We only considered the genes that were documented in the ligand-receptor interaction database, and further filtered out those detected in less than 20 % of the spots in one tumor tissue.

MERFISH ST Data

MERFISH data sets for four tumor samples (one tumor slice from one patient and three tumor slices from another patient) were downloaded from this website [37]. Unlike the 10X Genomics platform, MERFISH is a spatially resolved single-cell transcriptome profiling technology. For each tumor slice, before normalization, we filtered out the cells with less than 10 genes observed and filtered out the genes that were observed in less than 1500 cells. We then performed library-size normalization using the function *normalizeGiotto* implemented in `Giotto R` package. The average number of genes per cell ranged from 10 to 346 in the four tumor slices, while the number of cells after filtering were 355633, 248065, 68953 and 209173 in the four tumor slices, respectively. After pre-processing using the pipeline implemented under `Giotto` [19] including normalization and filtering of the cells with less then 10 genes measured and filtering of genes measured in less than 1500 cells, we clustered the cells of each tumor tissue using *doLeidenCluster* implemented under `Giotto`. We used the same single cell data as used for deconvolution of the spots in 10X ST data [30] to identify the marker genes upregulated in each of the 9 cell types mentioned above. We mapped the marker genes to the clusters in the MERFISH data, e.g., we combined the cells from all the clusters where the Epithelial marker genes were upregulated as defined the new class as tumor cluster, and so on. This way we identified the groups of cells as tumor, macrophage, T-cells, fibroblast or endothelial in each of the four tumor tissues. The processed data with cell-type labeling was then used for LR network construction and validation.

To identify the cells from different cell types that are closest to the tumor cells, we constructed a Delaunay network using the function *createSpatialNetwork* implemented under `Giotto`. For each tumor cell, we identified the cell adjacent/neighboring to the tumor cell that has minimum distance from the tumor cell, obtained from the network.

As in the 10X genomics ST data, for the network learning in MERFISH data also, we only considered the genes that were documented in the ligand-receptor interaction database, and further filtered out those detected in less than 10 % of the cells in one tumor tissue. The median of the total UMIs per cell in each sample was used for the downsampling normalization in `LRnetST`. Table A.3 in the Additional file 1 gives the detailed numbers of the tumor and neighboring cells, as well as the numbers of adjacent index-neighbor-cell pairs in each tumor tissue.

Supplementary Information

Additional file 1: This file contains supplementary methods description and tables.

Declaration

Ethics approval and consent to participate

Not applicable.

Author's contributions

Writing: SC, PY, WW, JP and PW. Data analysis and Visualization: SC, JP, PW, WW, SFB, PY, APC and SCM. Result interpretation: SC, JP, PW, WW, SFB, PY, APC and SCM. Supervision: WW, SCM, JP and PW. Funding Acquisition: WW and PW.

Acknowledgement
This work is supported by grants (U24CA271114, U01CA271407) and a grant (R01CA268380) from the National Cancer Institute. We also thank Dr. Guocheng Yuan and his team members for their instruction on the use of Giotto.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
An R package of `LRnetST` is made available as a github repository https://github.com/jie108/LRnetST. All the codes for reproducing the results along with the data will be made available through this github repository.

Author details
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Gynecologic Oncology and Reproductive Medicine, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [4]Department of Statistics, University of California Davis, Davis, CA, USA. [5]Department of Statistics, Rice University, Houston, TX, USA. [6]co-correspondence,.

References

1. Network, C.G.A.R., *et al.*: Integrated genomic analyses of ovarian carcinoma. Nature **474**, 609–615 (2011)
2. Patch, A.M., *et al.*: Whole-genome characterization of chemo-resistant ovarian cancer. Nature **521**(7553), 489–494 (2015)
3. Tothill, R.W., *et al.*: Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin. Cancer Res. **14**, 5198–5208 (2008)
4. Zhang, H., *et al.*: Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell **166**(3), 755–765 (2016)
5. Chowdhury, S., Kennedy, J.J., Ivey, I.G., Murillo, O.D., Hosseini, N., *et al.*: Proteogenomic analysis of chemo-refractory high-grade serous ovarian cancer. Cell **186**(16), 3476–3498 (2023)
6. Zhang, B., *et al.*: Revisiting ovarian cancer microenvironment: a friend or a foe? Protein Cell **9**(8), 674–692 (2018)
7. Tsujikawa, T., Mitsuda, J., Ogi, H., Miyagawa-Hayashino, A., Konishi, E., Itoh, K., Hirano, S.: Prognostic significance of spatial immune profiles in human solid cancers. Cancer Sci **111**(10), 3426–3434 (2020)
8. Efremova, M., Vento-Tormo, M., Teichmann, S.A., Vento-Tormo, R.: Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature protocols **15**(4), 1484–1506 (2020)
9. Cabello-Aguilar, S., et al.: Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Research **48**(10) (2020)
10. Jin, S., *et al.*: Inference and analysis of cell-cell communication using cellchat. Nature communications **12**(1), 1–20 (2021)
11. Noel, F., et al.: Dissection of intercellular communication using the transcriptome-based framework icellnet. Nature communications **12**(1089) (2021)
12. Nagai, J.S., *et al.*: Crosstalker: analysis and visualization of ligand–receptor networks. Bioinformatics **37**(22), 4263–4265 (2021)
13. Browaeys, R., Saelens, W., Saeys, W.: Nichenet: modeling intercellular communication by linking ligands to target genes. Nature methods **17**(2), 159–162 (2020)
14. Cheng, J., Zhang, J., Wu, Z., Sun, X.: Inferring microenvironmental regulation of gene expression from single-cell rna sequencing data using scmlnet with an application to covid-19. Briefings in bioinformatics **22**(2), 988–1005 (2021)
15. Cheng, J., Zhang, J., Wu, Z., Sun, X.: Cytotalk: De novo construction of signal transduction networks using single-cell transcriptomic data. Science Advances **7**(16), 1356 (2021)
16. Asp, M., Bergenstråhle, J., Lundeberg, J.: Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays **42**(10) (2020)
17. Marx, V.: Method of the year: spatially resolved transcriptomics. Nature Methods **18**(1), 9–14 (2021)
18. Ståhl, P.L., *et al.*: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
19. Dries, R., Zhu, Q., Dong, R., Eng, C.H.L., Li, H., Liu, K., *et al.*: Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome biology **22**, 1–31 (2021)
20. Pham, D., Tan, X., Xu, J., Grice, L.F., Lam, P.Y., *et al.*: stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. BioRxiv **22**, 2020–2025 (2020)
21. Pearl, J.: Causality: Models, Reasoning and Inference vol. 29. Cambridge Univ Press, ??? (2000)
22. Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., Schadt, E.E.: Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature genetics **40**(7), 854–861 (2008)
23. Zhu, J., Sova, P., Xu, Q., Dombek, K.M., Xu, E.Y., Vu, H.e.a.: Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLoS Biology **10**(4) (2012)
24. Sung, W.H., Gong, C., Myun-Seok, C., Hua, Z.: Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. Journal of the American Statistical Association **111**(515), 1004–1019 (2016)

25. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of computational biology **7**(3-4), 601–620 (2000)
26. PeŐer, D., Regev, A., Elidan, G., Friedman, N.: Inferring subnetworks from perturbed expression profiles. Bioinformatics **17**(suppl 1), 215–224 (2001)
27. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science Signalling **308**(5721), 523 (2005)
28. Chowdhury, S., Wang, R., Yu, Q., Huntoon, C.J., Karnitz, L.M., Kaufmann, S.H., et al.: Dagbagm: learning directed acyclic graphs of mixed variables with an application to identify protein biomarkers for treatment response in ovarian cancer. BMC Bioinformatics **23** (2022)
29. Scutari, M.: Learning bayesian networks with the bnlearn r package. Journal of Statistical Software **35**(3) (2010)
30. Geistlinger, L., Oh, S., Ramos, M., Schiffer, L., LaRue, R.S., Henzler, C.M., et al.: Multiomic analysis of subtype evolution and heterogeneity in high-grade serous ovarian carcinoma. Cancer research **80**(20), 4335–4345 (2020)
31. Ramilowski, J., Goldberg, T., Harshbarger, J., *et al.*: A draft network of ligand–receptor-mediated multicellular signalling in human. Nat Commun **6**(7866), 755–765 (2015)
32. He, X., Zhang, J.: Why do hubs tend to be essential in protein networks? PLoS genetics **2**(6) (2006)
33. Zhou, W., Ma, J., Zhao, H., Wang, Q., Guo, X., Chen, L., et al.: Serum exosomes from epithelial ovarian cancer patients contain lrp1, which promotes the migration of epithelial ovarian cancer cell. Molecular Cellular Proteomics **22**(4) (2023)
34. Strickland, D.K., Muratoglu, S.C., Antalis, T.M.: Serpin–enzyme receptors: Ldl receptor-related protein 1. Methods in enzymology **499**, 17–31 (2011)
35. Miao, R., Dong, X., Gong, J., Li, Y., Guo, X., Wang, J.e.a.: Examining the development of chronic thromboembolic pulmonary hypertension at the single-cell level. Hypertension **79**(3), 562–574 (2022)
36. Wu, J., Chen, Z.P., Shang, A.Q., Wang, W.W., Chen, Z.N., Tao, Y.J., et al.: Systemic bioinformatics analysis of recurrent aphthous stomatitis gene expression profiles. Oncotarget **8**(67) (2017)
37. Vizgen merfish ffpe human immuno-oncology data set. Vizgen (2022)
38. Lillis, A.P., Van Duyn, L.B., Murphy-Ullrich, J.E., Strickland, D.K.: Ldl receptor-related protein 1: unique tissue-specific functions revealed by selective gene knockout studies. Physiological reviews **88**(3), 887–918 (2008)
39. Xing, P., Liao, Z., Ren, Z., Zhao, J., Song, F., Wang, G., et al.: Roles of low-density lipoprotein receptor-related protein 1 in tumors. Chinese journal of cancer **35**(1), 1–8 (2016)
40. Potere, N., Del Buono, M.G., Mauro, A.G., Abbate, A., Toldo, S.: Low density lipoprotein receptor-related protein-1 in cardiac inflammation and infarct healing. Frontiers in cardiovascular medicine **6**(51) (2019)
41. Kinny-Köster, B., Guinn, S., Tandurella, J.A., Mitchell, J.T., Sidiropoulos, D.N., Loth, M., et al.: Inflammatory signaling in pancreatic cancer transfers between a single-cell rna sequencing atlas and co-culture. bioRxiv (2022)
42. Zhang, X., Wang, R., Chen, H., Jin, C., Jin, Z., Lu, J., et al.: Aged microglia promote peripheral t cell infiltration by reprogramming the microenvironment of neurogenic niches. Immunity Ageing **19**(1) (2022)
43. Verma, T., Pearl, J.: Equivalence and synthesis of causal models. In Henrion, M., Shachter, R. Kanal, L., and Lemmer, J., editors, Proceeding of the Sixth Conference on Uncertainty in Artificial Intelligence, 220–227 (1991)
44. Geiger, D., Heckerman, D.: Learning gaussian networks. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, pp. 235–243 (1994). Morgan Kaufmann Publishers Inc.
45. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search vol. 81. MIT press, ??? (2001)
46. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. The Journal of Machine Learning Research **8**, 613–636 (2007)
47. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine learning **65**(1), 31–78 (2006)
48. Ferri-Borgogno, S., Zhu, Y., Sheng, J., Burks, J.K., Gomez, J.A., Wong, K.K., *et al.*: Spatial transcriptomics depict ligand–receptor cross-talk heterogeneity at the tumor-stroma interface in long-term ovarian cancer survivors. Cancer Research **83**(9), 1503–1516 (2023)
49. Navarro, J.F., *et al.*: St pipeline: an automated pipeline for spatial mapping of unique transcripts. Bioinformatics **33**(16), 2591–2593 (2017)
50. Dobin, A., *et al.*: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)

Figure Legends

Main

Figure 1: `LRnetST` pipeline.

(A) Main steps: Given the ST data matrix and the spatial information, `LRnetST` first generates multiple downsampled data matrices, and a list of matching Index Spot and Neighbor Spot pairs by identifying adjacent spot pairs enriched with tumor and non-tumor cells, respectively. Then, from each downsampled data matrix, a neighbor integrated matrix (NIM) is constructed and a bootstrap resample of the NIM is generated. DAGs are learned from these bootstrap resamples and are aggregated using structural hamming distance.
(B) Construction of one DAG: For each down-sampled data matrix Z, `LRnetST` first derives an initial NIM by stacking the gene expression profiles of the matching Index and Neighbor Spots. Then, the downsample-normalized UMI counts Z are converted into activation status Y and "expression level" X given Y, which leads to the final NIM. A bootstrap resample is then generated by sampling without replacement of the columns of this NIM and a DAG is learned based on the resample.

Figure 2: Reproducibility of the edges detected in the network learned on the 10X ST data.

(A) Common edges inferred by LRnetST in the 4 10X ST data.
(B) Common edges inferred by SpatCellCellCom (Giotto) in the 4 10X ST data.
(C) Common edges inferred by stLearn in the 4 10X ST data.
(D) P-values from the fisher exact tests assessing whether the overlapping between the LR interactions of a pair of tumors is significant by LRnetST, SpatCellCellCom (Giotto) and stLearn.
(E) ITGB1 module by LRnetST in 10X ST data sample 10, where ITGB1 in the tumor spot is causally associated with several genes from the neighboring stroma spot.
(F) ITGB1 module by LRnetST in 10X ST data sample 10, where ITGB1 in the neighboring stroma spot is causally associated with several genes from the tumor spot.
(G) LRP1 module by LRnetST in 10X ST data sample 10, where LRP1 in the tumor spot is causally associated with several genes from the neighbor spot.
(H) LRP1 module by LRnetST in 10X ST data sample 10, where LRP1 in the neighboring stroma spot is causally associated with several genes from the tumor spot.

Figure 3: Ligand receptor network learning by LRnetST considering index and different types of neighboring cells in MERFISH data of HGSC patient 2 slice 2.
(A) Spatial plots showing the index (tumor) cells in red and the neighboring machrophage cells close to the tumor cells in blue, respectively. In the background, all the epithelial/tumor cells are indicated by pink and all the macrophage cells are highlighted by light blue, while the remaining cells are plotted as light yellow.
(B) – (E) Ligand-receptor network topology considering tumor cells and different types of neighboring cells: immune, macrophage, fibroblast and endothelial, respectively. The edges highlighted in red are also inferred in the LR network learned based on the 10X ST data. Note in these sub-networks, LR interactions inferred in at least 2 MERFISH data sets are shown.

Figure 4: Ligand receptor (LR) network topology showing edges inferred from 10X ST data in at least 2 out of 4 samples.
(A) LR network topology of the edges inferred from 10X ST data in at least 2 out of 4 samples by LRnetST. The edges highlighted in red are inferred only in the two short term survivor patients and are validated based on MERFISH data in at least one of the 4 HGSC samples, while the edges highlighted in green are inferred in the mixture of long- and short-term survivor patients and are validated based on MERFISH data in at least one of the 4 HGSC samples. Note that, we only show the genes that appear in the module where at least one reproducible edge inferred from the 10X data was validated in the MERFISH data.
(B) LR network topology of the edges inferred from 10X ST data in at least 2 out of 4 samples by stLearn. The edges highlighted in red are also validated based on MERFISH data in at least one of the 4 HGSC samples.
(C) LR network topology of the edges inferred from 10X ST data in at least 2 out of 4 samples by SpatCellCellCom (Giotto). The edges highlighted in red are also validated based on MERFISH data in at least one of the 4 HGSC samples.

Supplementary Figure Legends

Figure S1: Impact of the number of genes (p) on the performance of different methods when sample size n = 400.
(A-F) True DAGs with $p = 100$, 200, 300, 400, 500, 600 and number of edges $|\mathbb{E}| = 80$, 160, 240, 320, 400, 480, respectively.
(G) F1 scores for detecting the skeleton edges by the four methods.
(H) F1 scores for detecting the directed edges by the four methods.

Figure S2: HE image for sample A4, A5, A10 and A12 and the ST grids (spots) mapped to the tissue slice.

Figure S3: Top hub nodes in the ligand-receptor network by LRnetST, SpatCellCellCom and stLearn.
(A) Barplots showing the number of parent nodes, child nodes, and total edges of each hub node for the top 30 hubs in the LR network by LRnetST.
(B) Barplots showing the total total edges of each hub node for the top 30 hubs in the LR network by SpatCellCellCom (Giotto).
(C) Barplots showing the total edges of each hub node for the top 30 hubs in the LR network by stLearn.

Figure S4: Ligand-receptor network by LRnetST based on 10X ST data of the two long term survivor samples (A4 and A5).
(A) Ligand-receptor network learned by LRnetST for sample A4.
(B) Ligand-receptor network learned by LRnetST for sample A5.

Figure S5: Ligand-receptor network by LRnetST based on 10X ST data of the two short term survivor samples (A10 and A12).
(A) Ligand-receptor network learned by LRnetST for sample A10.
(B) Ligand-receptor network learned by LRnetST for sample A12.

Figure S6: LRP1 module in the ligand-receptor network inferred by LRnetST based on 10X ST data of sample A4, A5 and A12. Note, the results for sample A10 are shown in Figure 2.

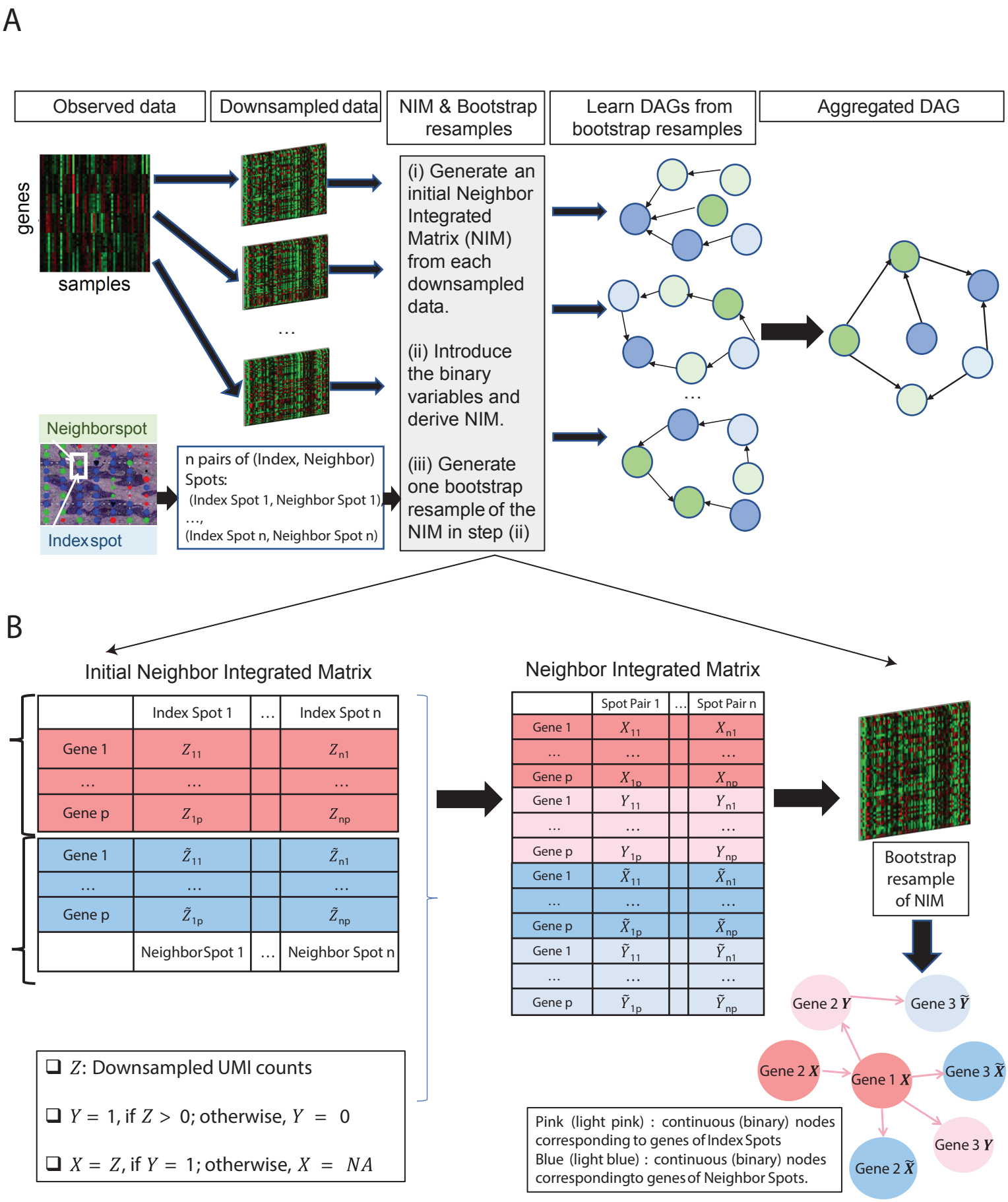
(A, C, E) LRP1 module by `LRnetST` in 10X ST data for sample A4, A5 and A12, where LRP1 in the tumor spot is causally associated with several genes from the neighbor spot.
(B, D, F) LRP1 module by `LRnetST` in 10X ST data sample A4, A5 and A12, where LRP1 in the neighboring stroma spot is causally associated with several genes from the tumor spot.

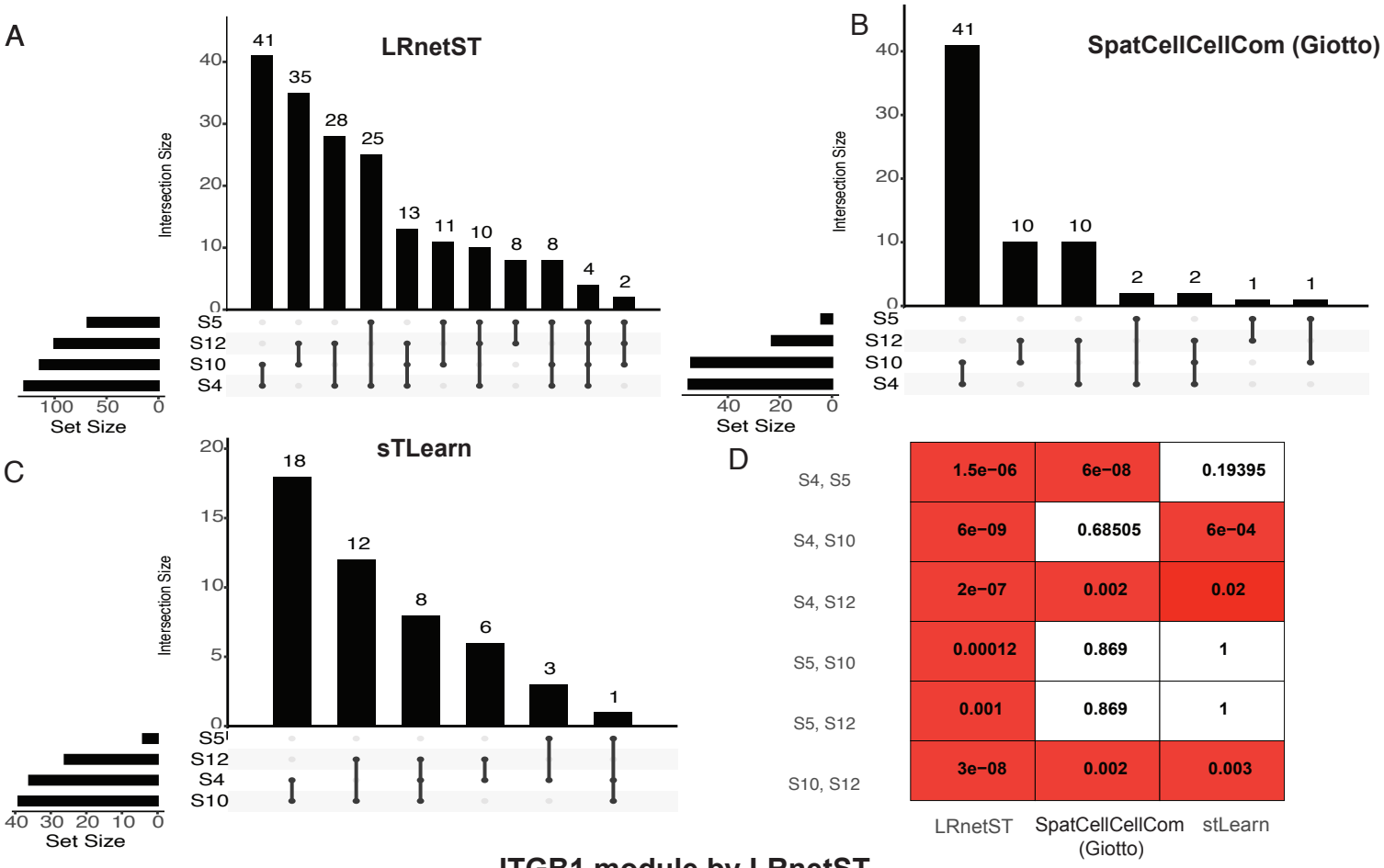
Figure S7: ITGB1 module in the ligand-receptor network inferred by `LRnetST` based on 10X ST data for samples A4, A5 and A12. Note, the results for sample A10 are shown in Figure 2.
(A, B, C) ITGB1 module by `LRnetST` on 10X ST data for samples A4, A5 and A12, where ITGB1 in the tumor spot is causally associated with several genes from the neighbor spot.
(D) ITGB1 module by `LRnetST` in 10X ST data sample A12, where ITGB1 in the neighboring stroma spot is causally associated with several genes from the tumor spot.

Figure S8: Spatial plots showing the index and different types of neighboring cells in MERFISH data of ovarian cancer patient 2 slice 2.
(A) Spatial plots showing the index (tumor) cells in red and the neighboring immune cells close to the tumor cells in blue, respectively. In the background, all the epithelial/tumor cells are indicated by pink and all the immune cells are highlighted by light blue, while the remaining cells are plotted as light yellow.
(B) Spatial plots showing the index (tumor) cells in red and the neighboring fibroblast cells close to the tumor cells in blue, respectively. In the background, all the epithelial/tumor cells are indicated by pink and all the fibroblast cells are highlighted by light blue, while the remaining cells are plotted as light yellow.
(C) Spatial plots showing the index (tumor) cells in red and the neighboring endothelial cells close to the tumor cells in blue, respectively. In the background, all the epithelial/tumor cells are indicated by pink and all the endothelial cells are highlighted by light blue, while the remaining cells are plotted as light yellow.

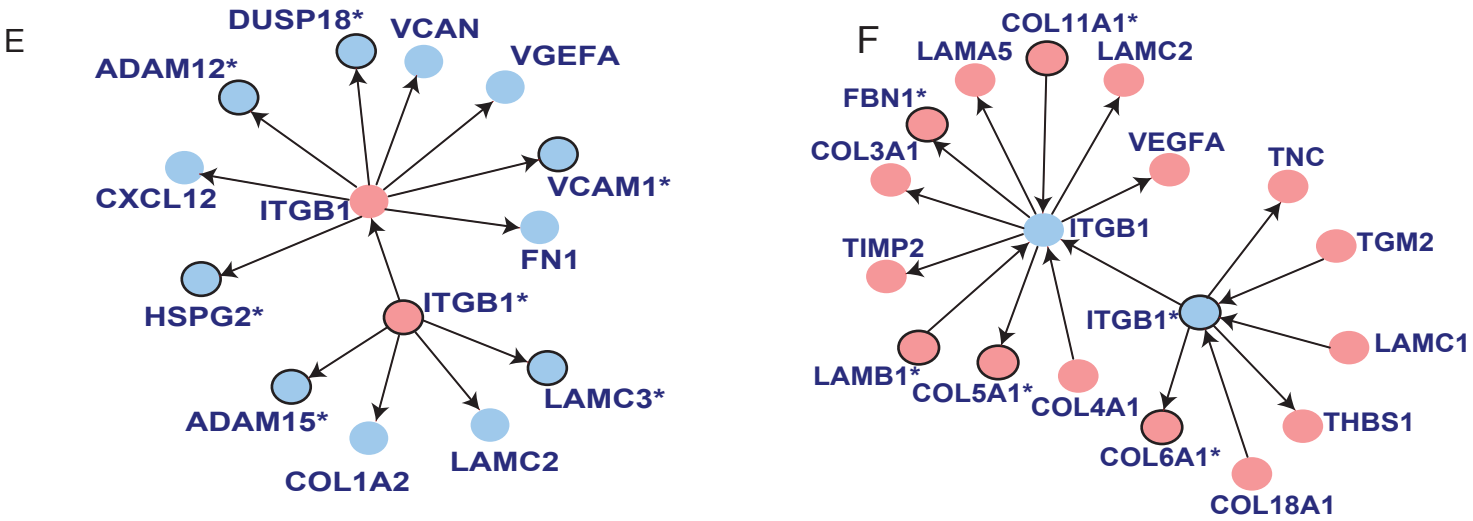
Figure S9: Clustering and cell type annotation results for MERFISH data of ovarian cancer patient 2 slice 2.
(A) UMAP plot showing the leiden clustering result.
(B) Heatmap showing the average gene expressions across 12 clusters for the cell-type specific marker genes.

LRnetST Pipeline

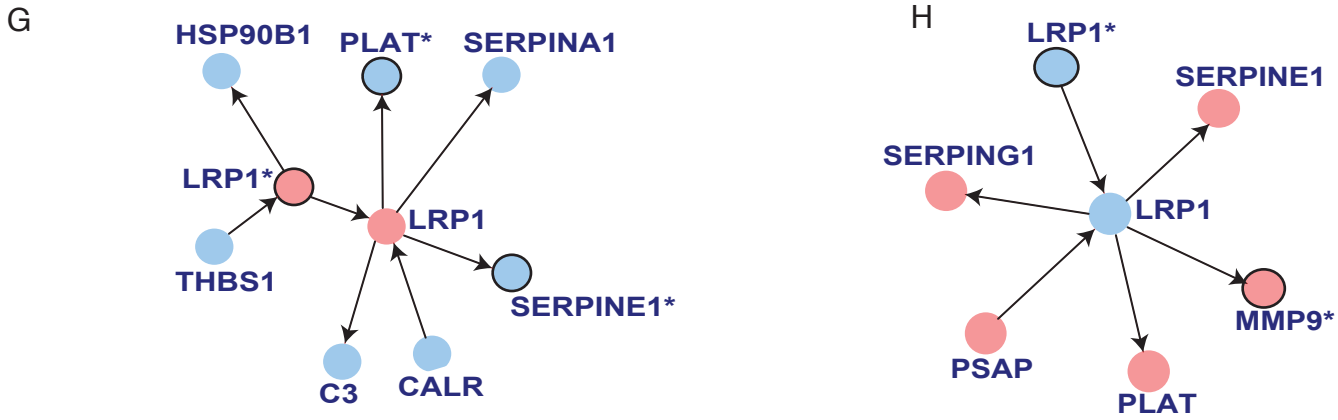




ITGB1 module by LRnetST in 10x genomics data for sample A10



LRP1 module by LRnetST in 10x genomics data for sample A10

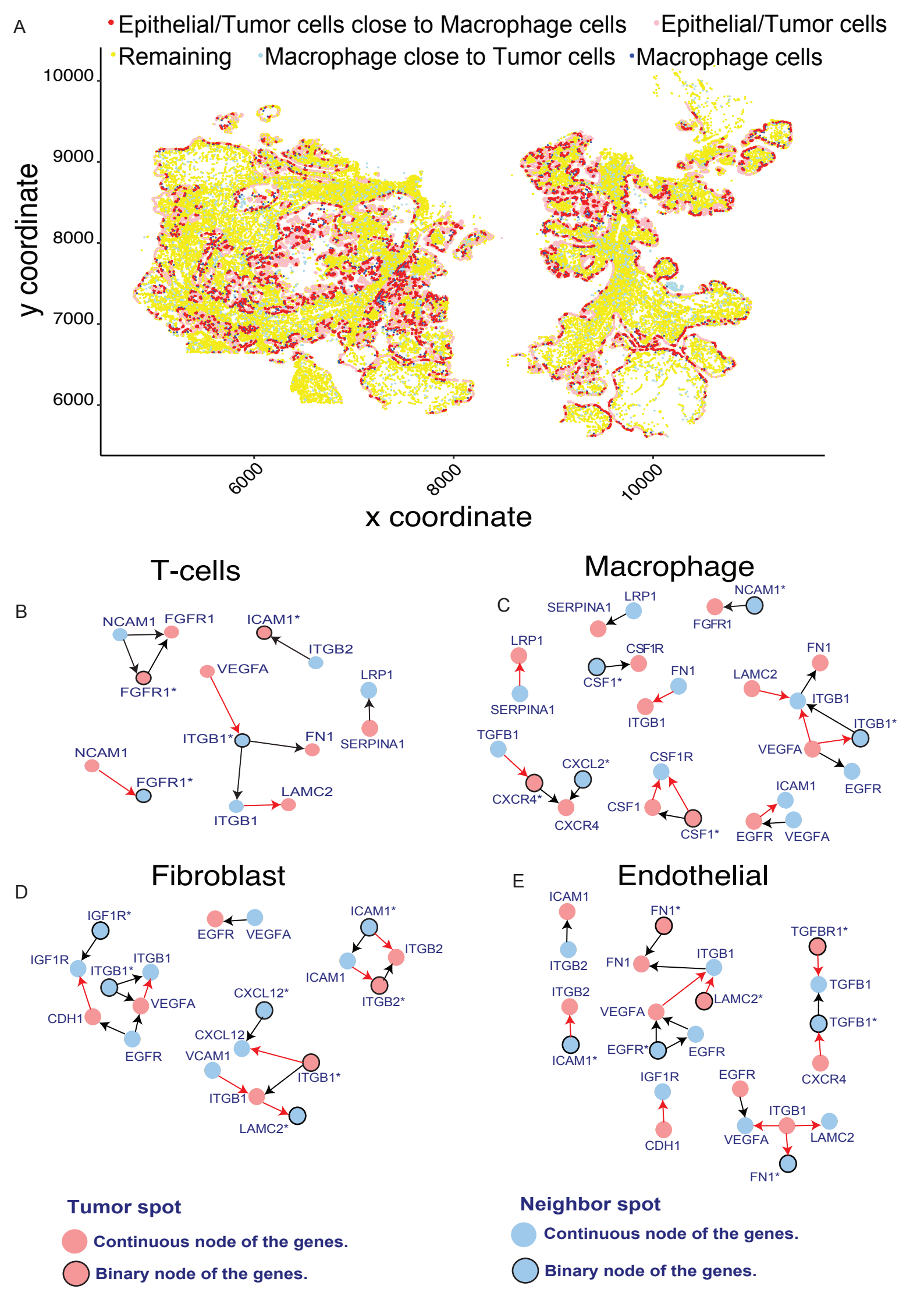


Tumor spot

- Continuous node of the genes.
- Binary node of the genes.

Neighbor spot

- Continuous node of the genes.
- Binary node of the genes.

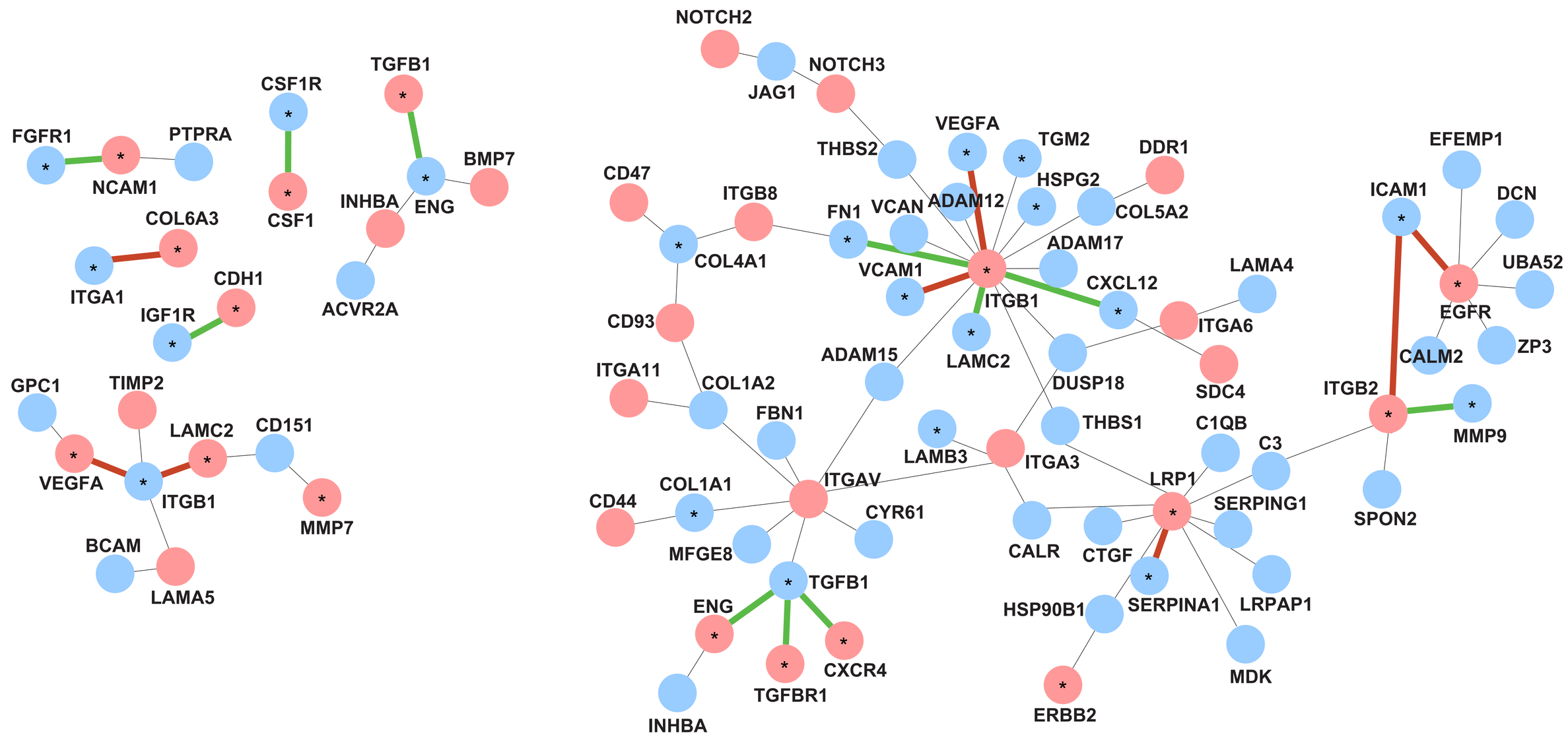

A
Epithelial/Tumor cells close to Macrophage cells
Epithelial/Tumor cells
Remaining
Macrophage close to Tumor cells
Macrophage cells
y coordinate
x coordinate
10000
9000
8000
7000
6000
6000
8000
10000
T-cells
B
NCAM1
FGFR1
FGFR1*
ICAM1*
ITGB2
VEGFA
LRP1
SERPINA1
ITGB1*
FN1
NCAM1
FGFR1*
LAMC2
ITGB1
Macrophage
C
SERPINA1
LRP1
NCAM1*
FGFR1
LRP1
CSF1R
CSF1*
FN1
SERPINA1
ITGB1
FN1
LAMC2
ITGB1
ITGB1*
TGFB1
VEGFA
CXCL2*
CSF1R
EGFR
CXCR4*
ICAM1
CSF1
CXCR4
CSF1*
EGFR
VEGFA
Fibroblast
D
IGF1R*
EGFR
VEGFA
ICAM1*
IGF1R
ITGB1
ITGB2
ITGB1*
ICAM1
VEGFA
CXCL12*
ITGB2*
CDH1
CXCL12
EGFR
VCAM1
ITGB1*
ITGB1
LAMC2*
Endothelial
E
ICAM1
FN1*
TGFBR1*
ITGB2
FN1
ITGB1
TGFB1
ITGB2
LAMC2*
VEGFA
TGFB1*
EGFR*
EGFR
ICAM1*
CXCR4
IGF1R
EGFR
ITGB1
VEGFA
LAMC2
CDH1
FN1*
Tumor spot
Continuous node of the genes.
Binary node of the genes.
Neighbor spot
Continuous node of the genes.
Binary node of the genes.

A

LRnetST

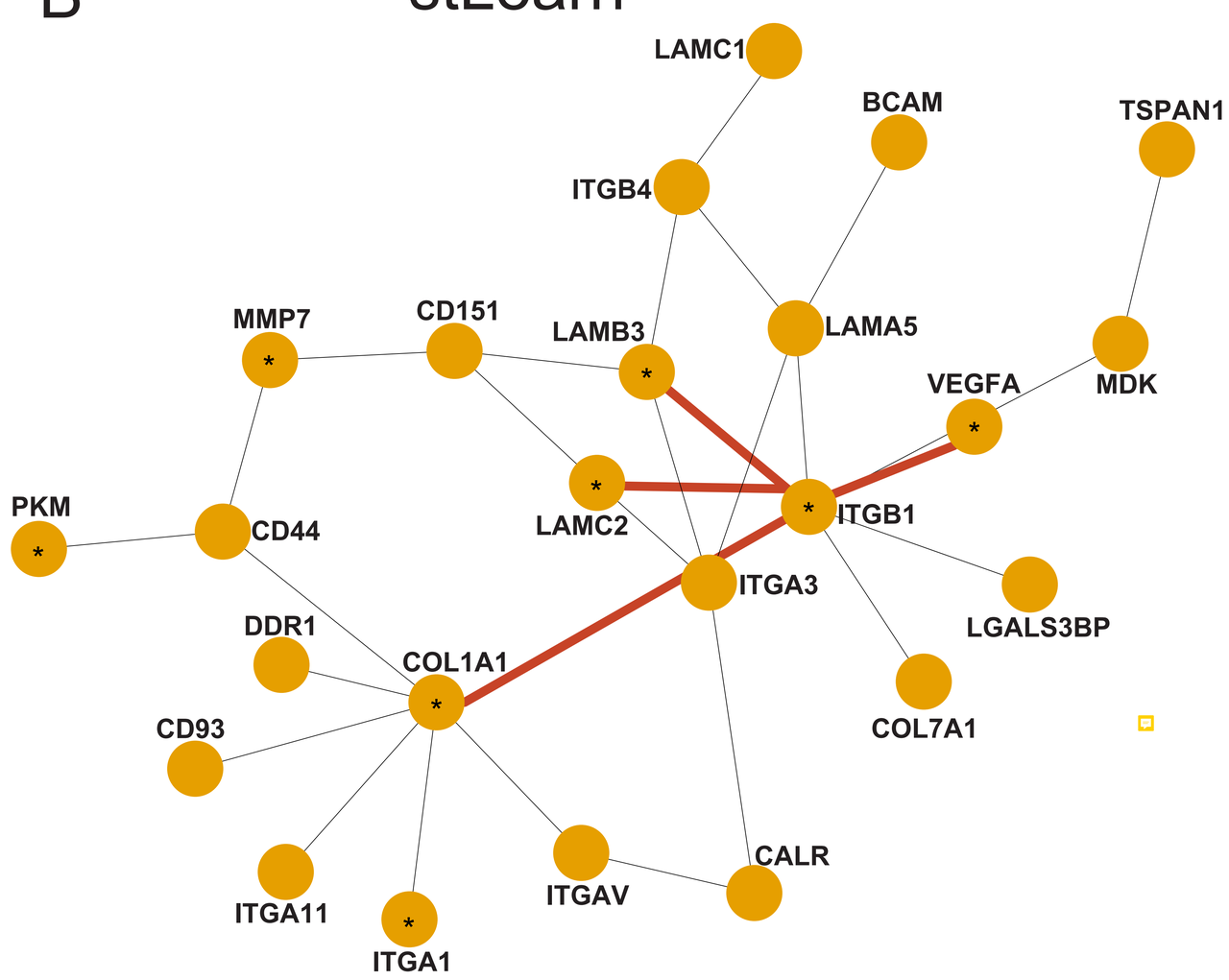
Edges inferred in the 10x genomics data of short term survivor samples

Edges inferred in the 10x genomics data of both short and long term survivor samples



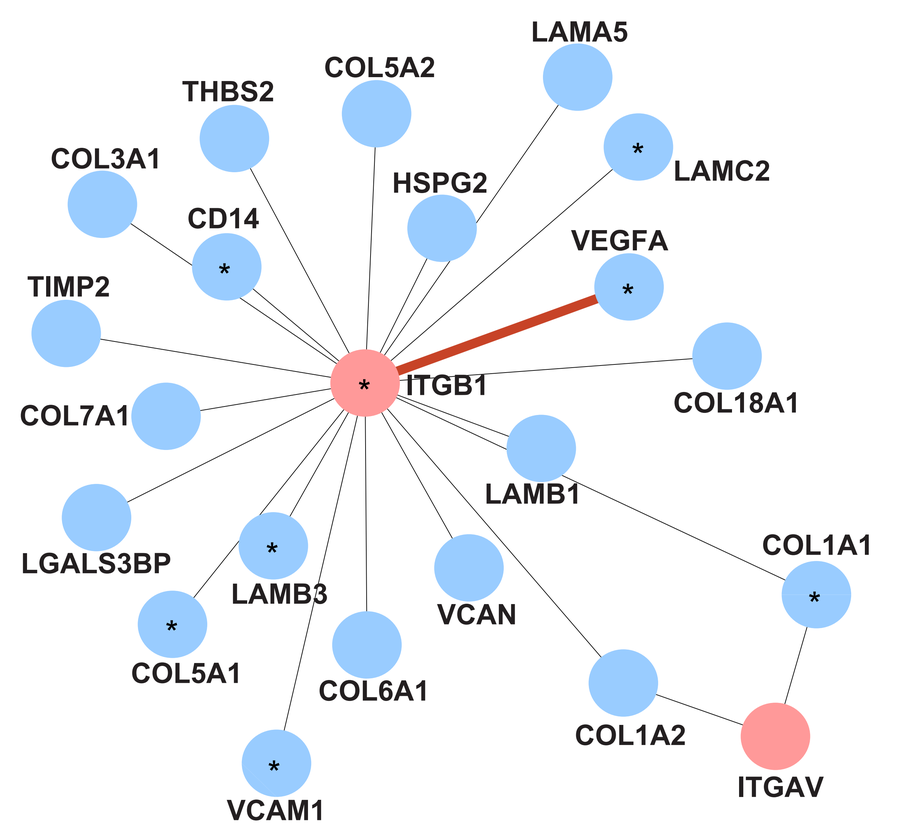
B

stLearn



C

SpatCellCellCom (Giotto)

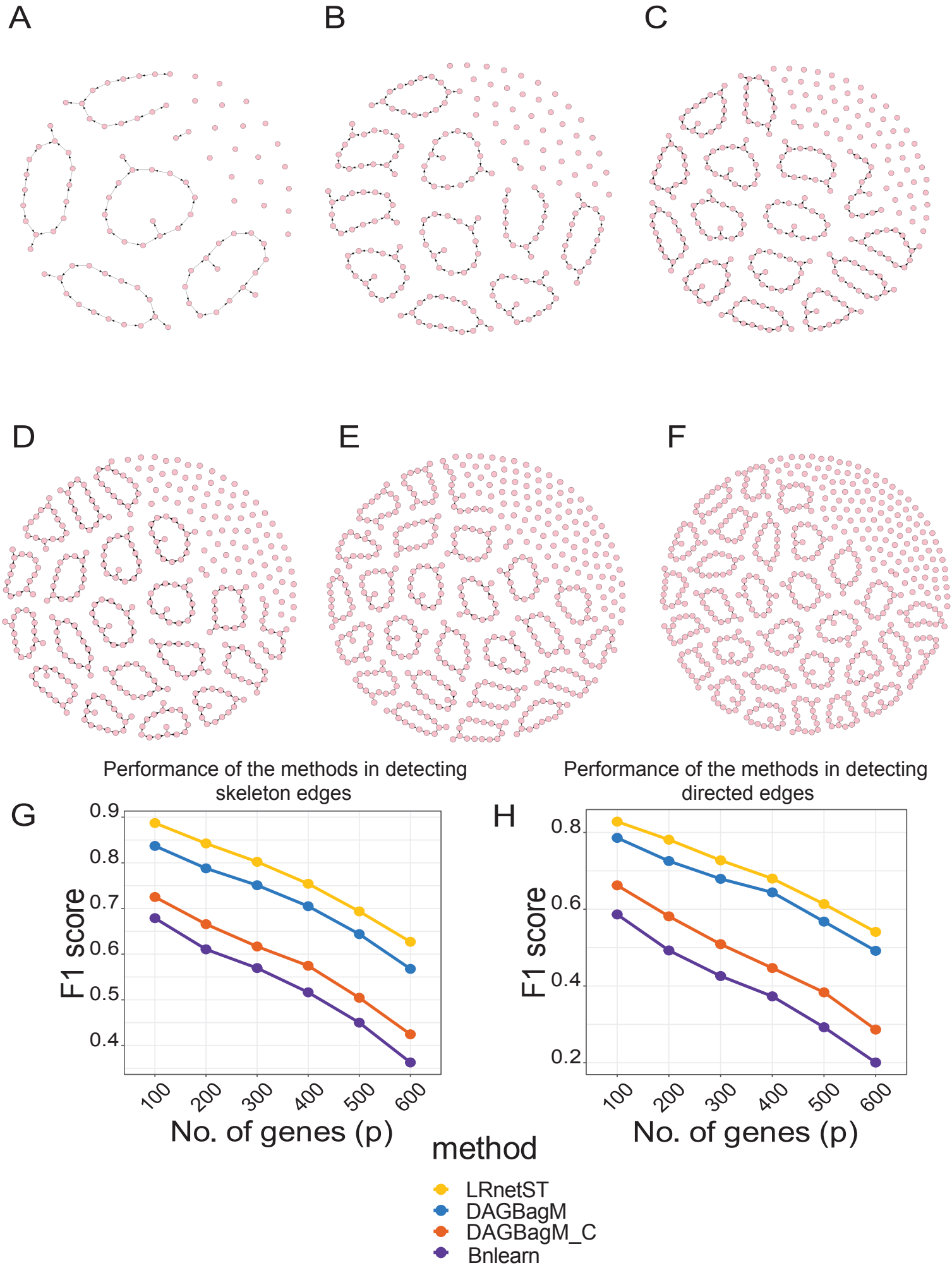


Genes expressed in the tumor spot.

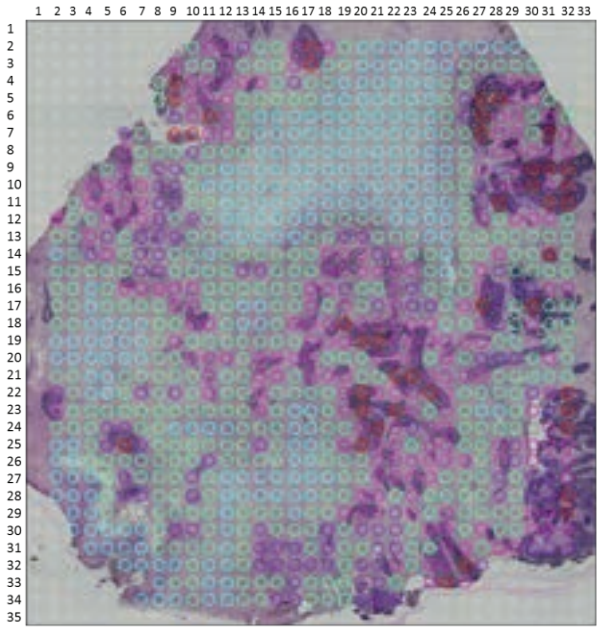
Genes expressed in the neighbor spot.

Nodes with * indicates that those genes are observed in both MERFISH and 10X ST.

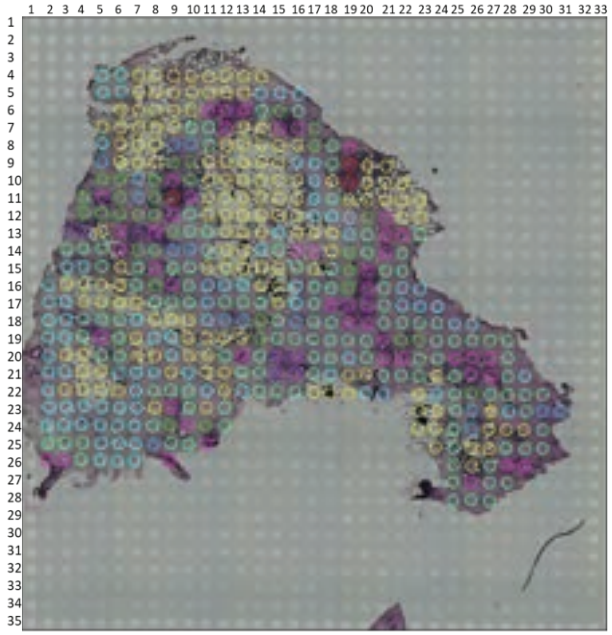
Genes expressed in tumor and/or neighbor spot.


A
B
C
D
E
F
G
Performance of the methods in detecting skeleton edges
F1 score
0.9
0.8
0.7
0.6
0.5
0.4
100
200
300
400
500
600
No. of genes (p)
H
Performance of the methods in detecting directed edges
F1 score
0.8
0.6
0.4
0.2
100
200
300
400
500
600
No. of genes (p)
method
LRnetST
DAGBagM
DAGBagM_C
Bnlearn

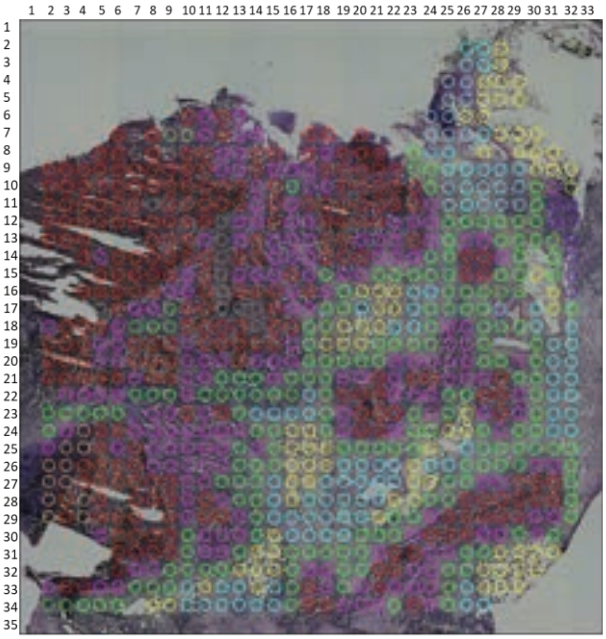
Sample A4



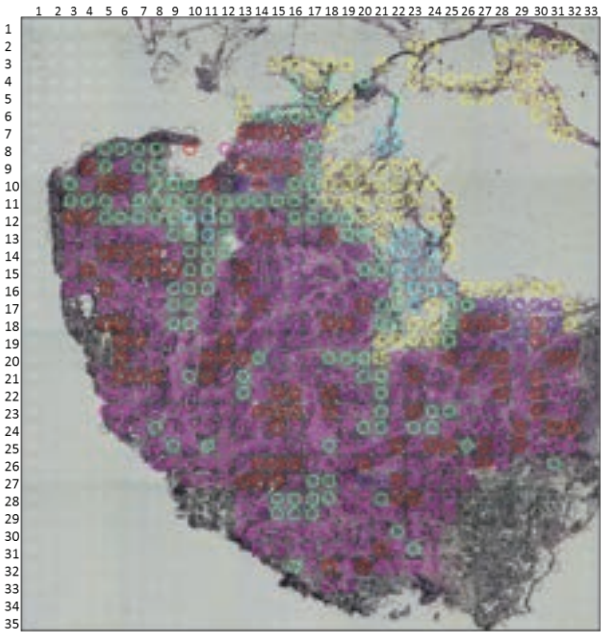
Sample A5

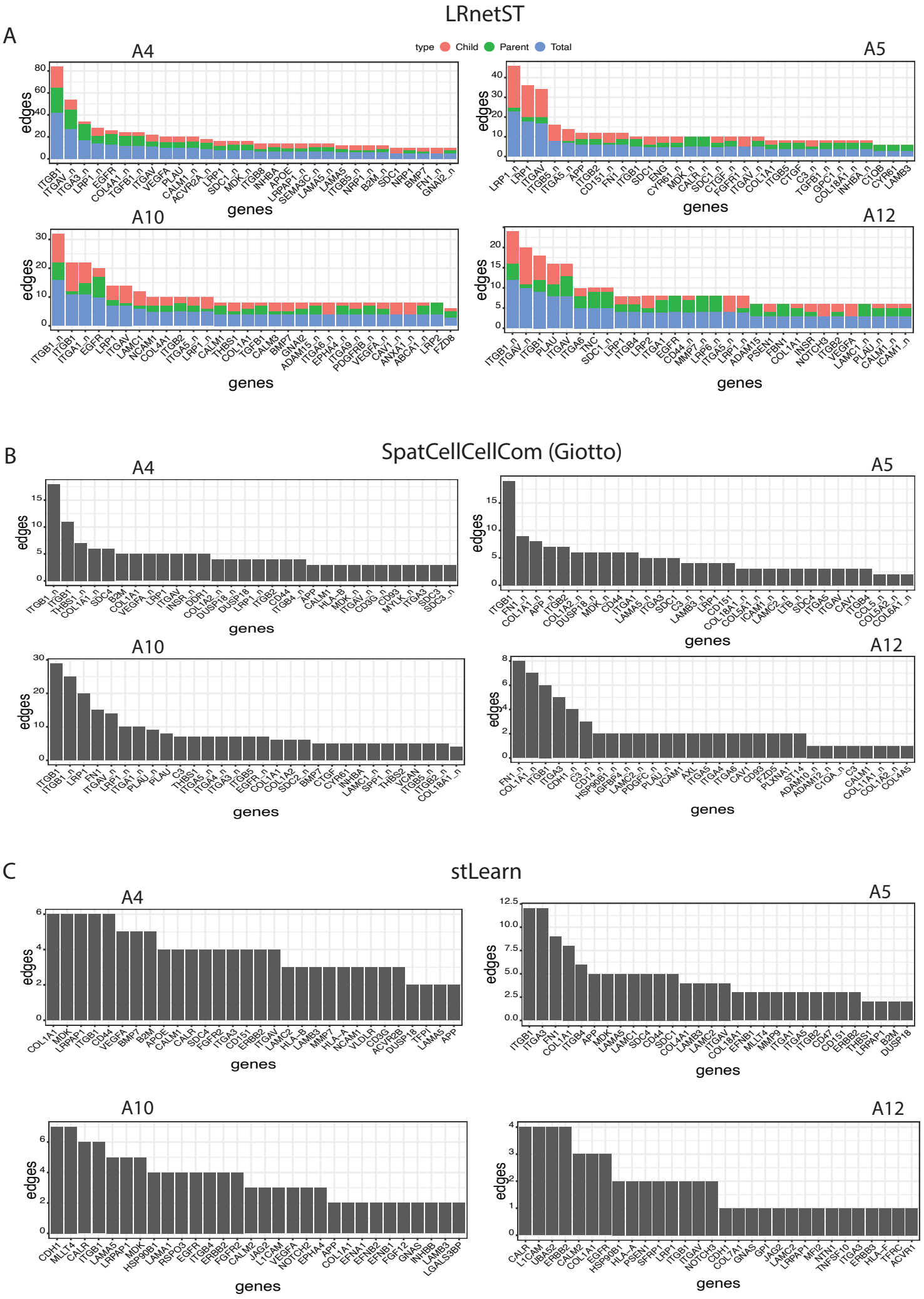


Sample A10



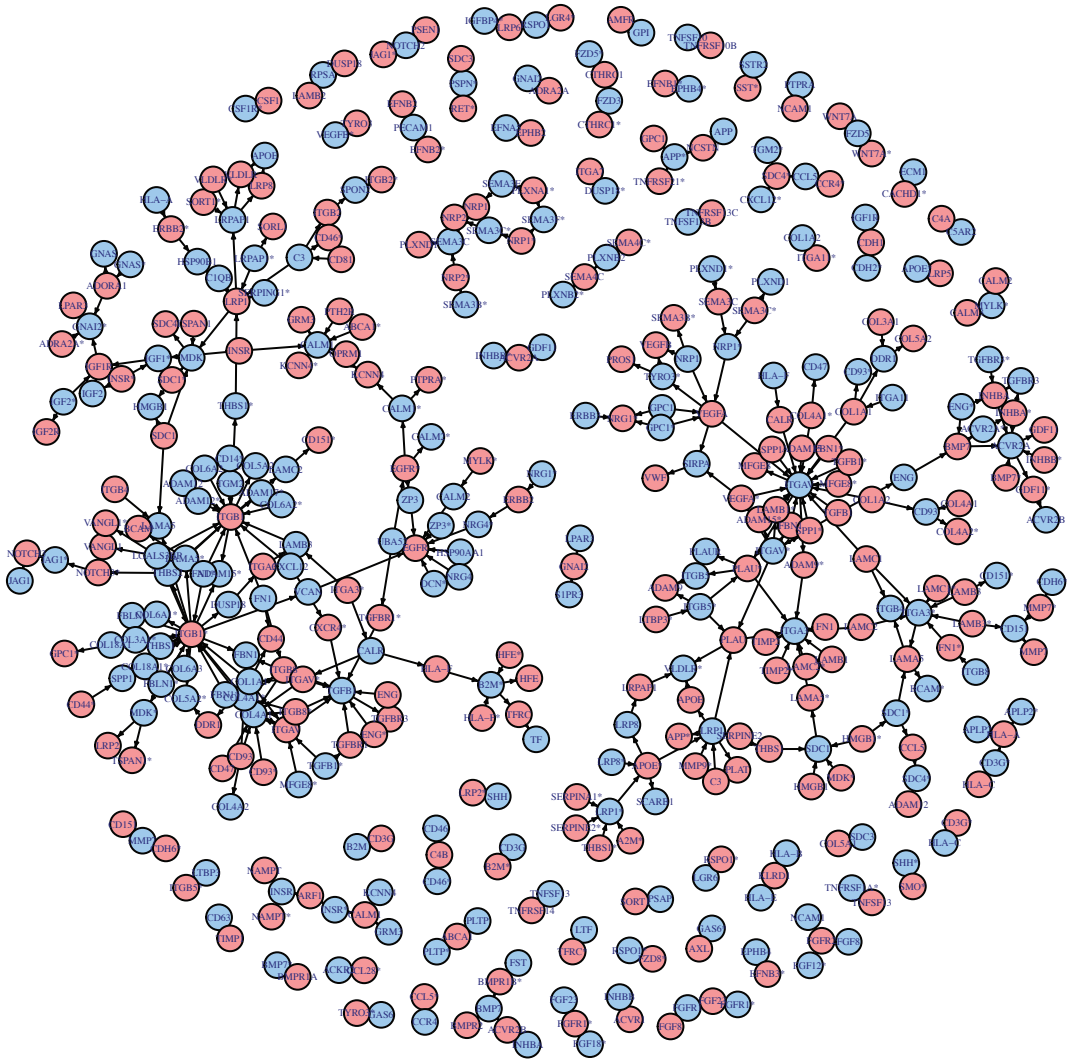
Sample A12




LRnetST
A
A4
type Child Parent Total
A5
A10
A12
edges
genes
B
SpatCellCellCom (Giotto)
A4
A5
A10
A12
edges
genes
C
stLearn
A4
A5
A10
A12
edges
genes

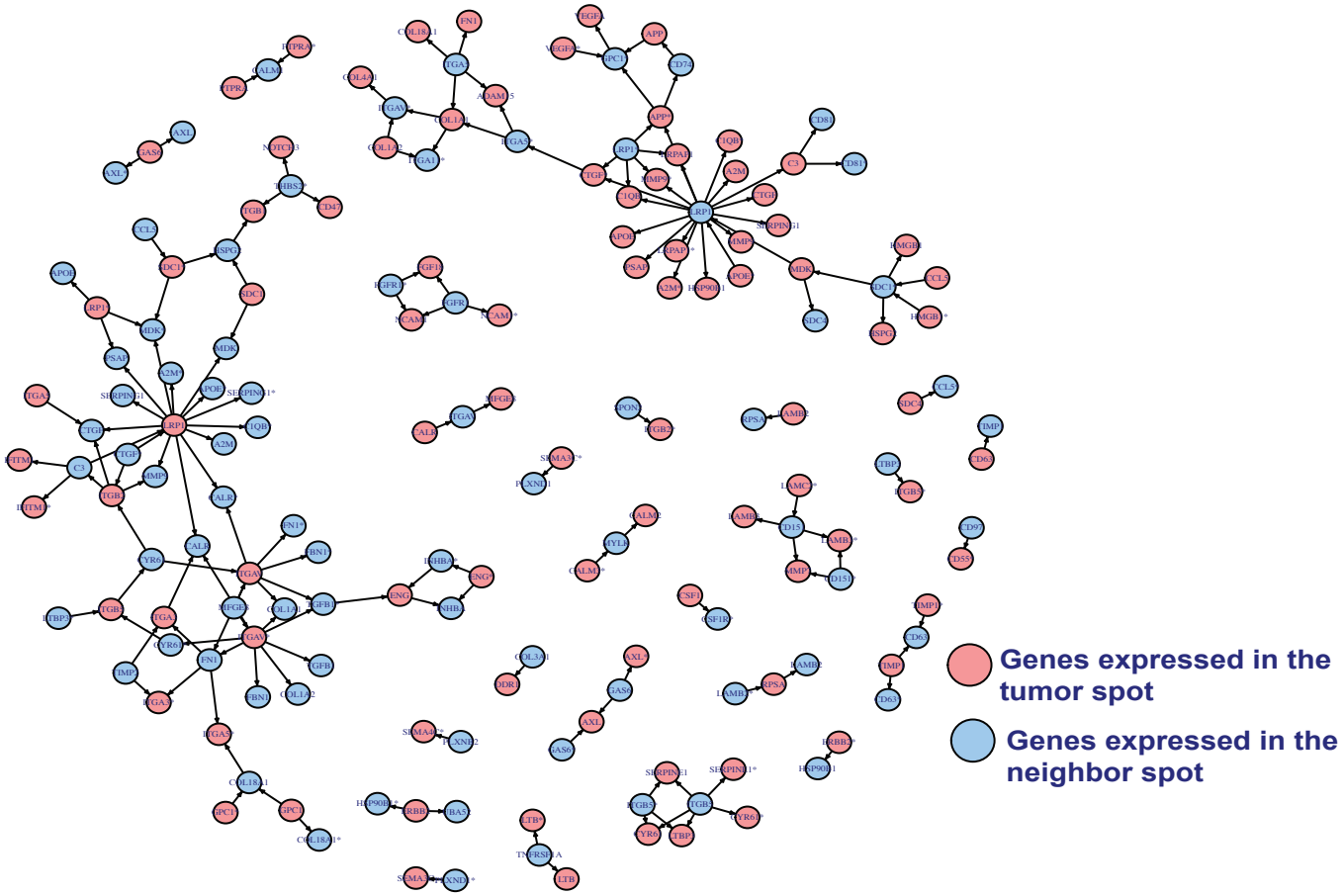
A

Inferred edges by LRnetST in sample A4



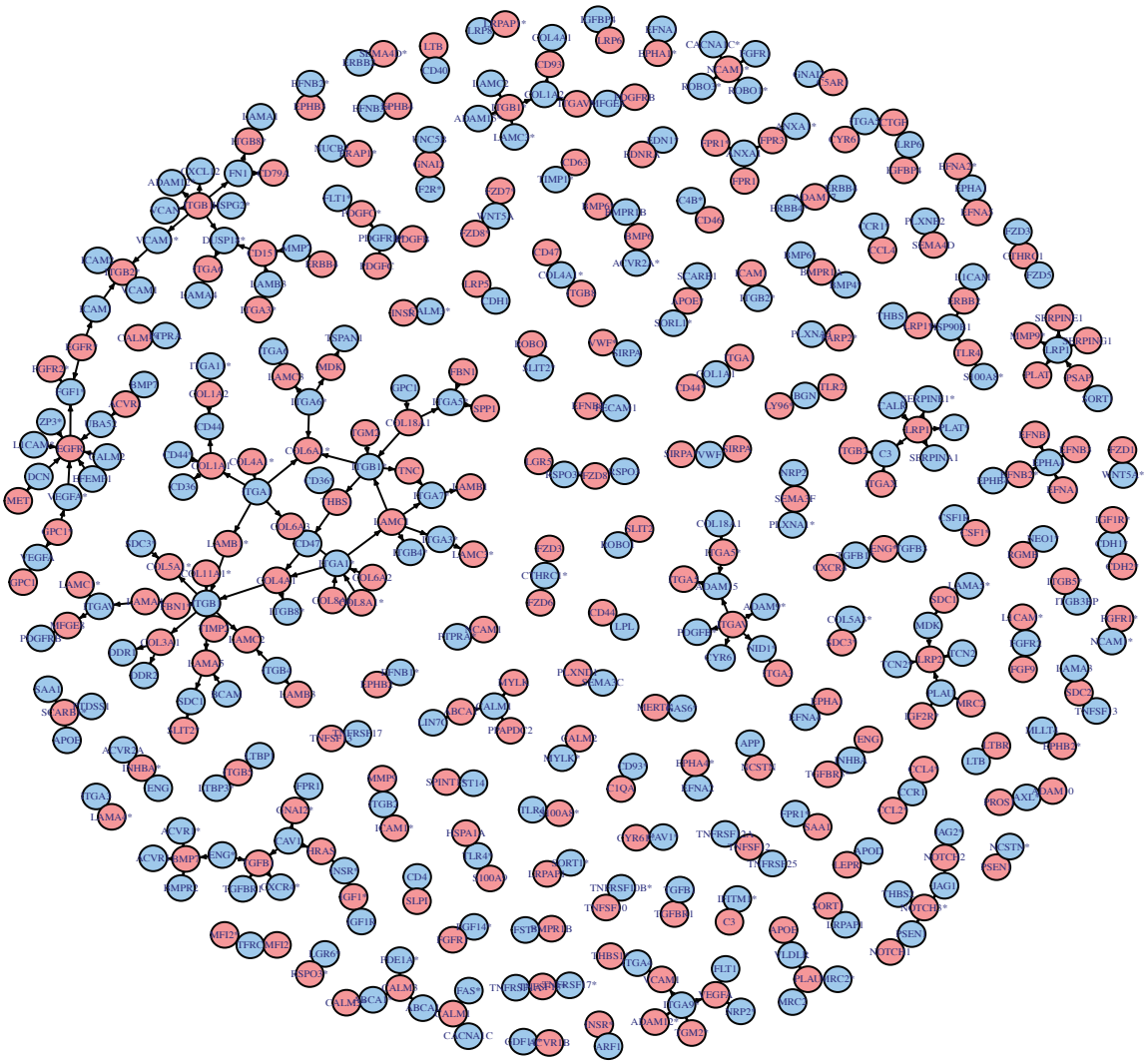
B

Inferred edges by LRnetST in sample A5



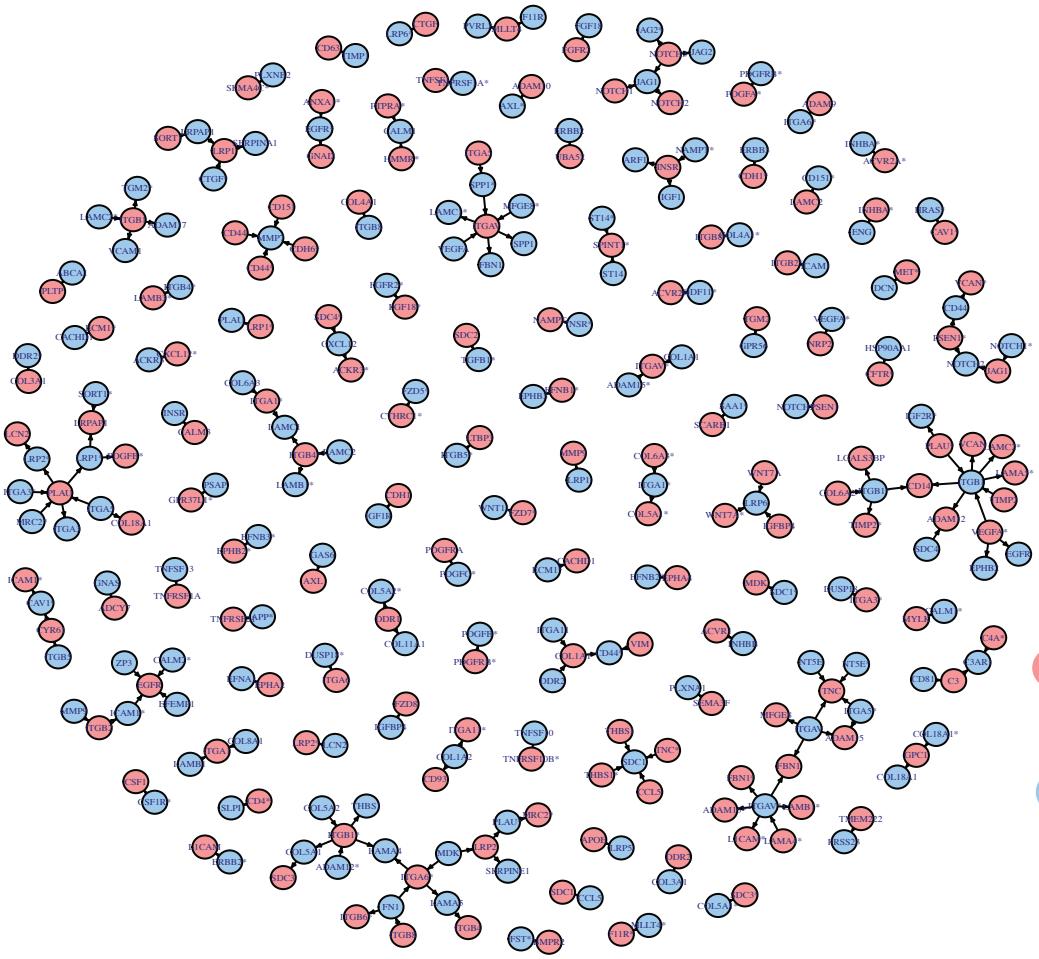
A

Inferred edges by LRnetST in sample A10



B

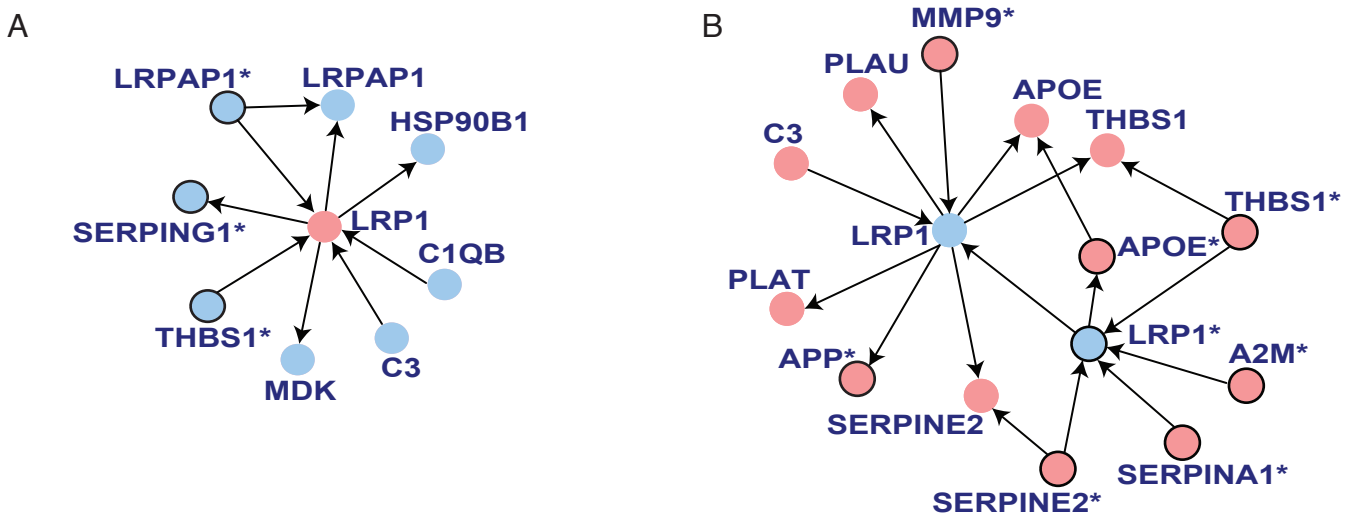
Inferred edges by LRnetST in sample A12



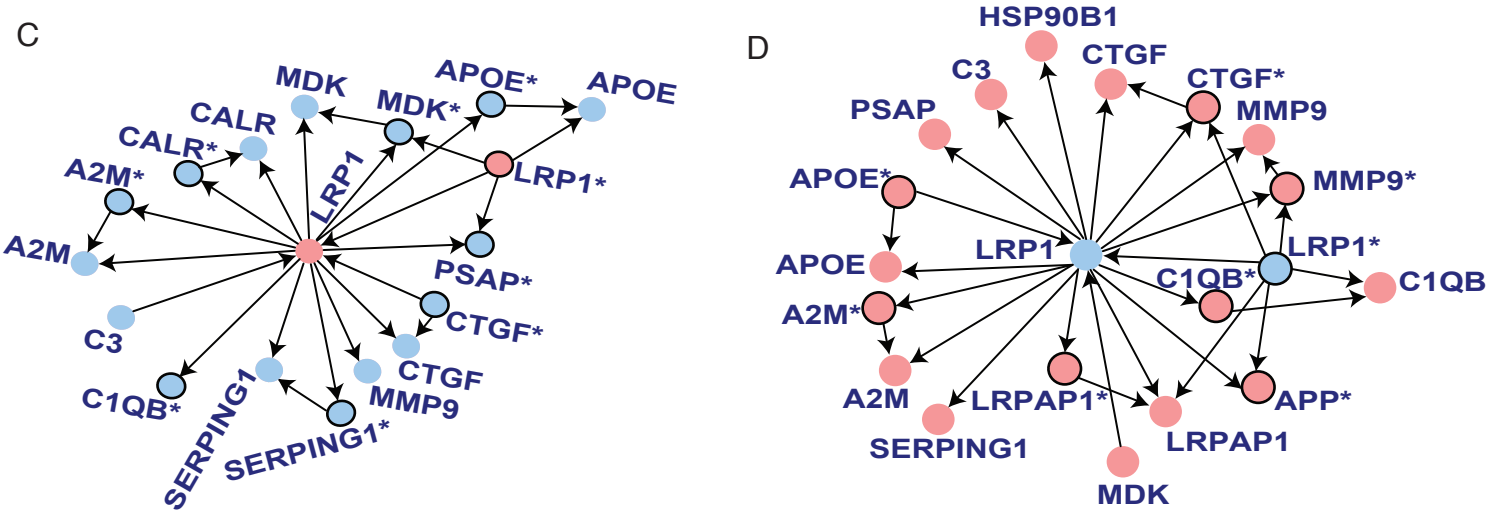
Genes expressed in the tumor spot

Genes expressed in the neighbor spot

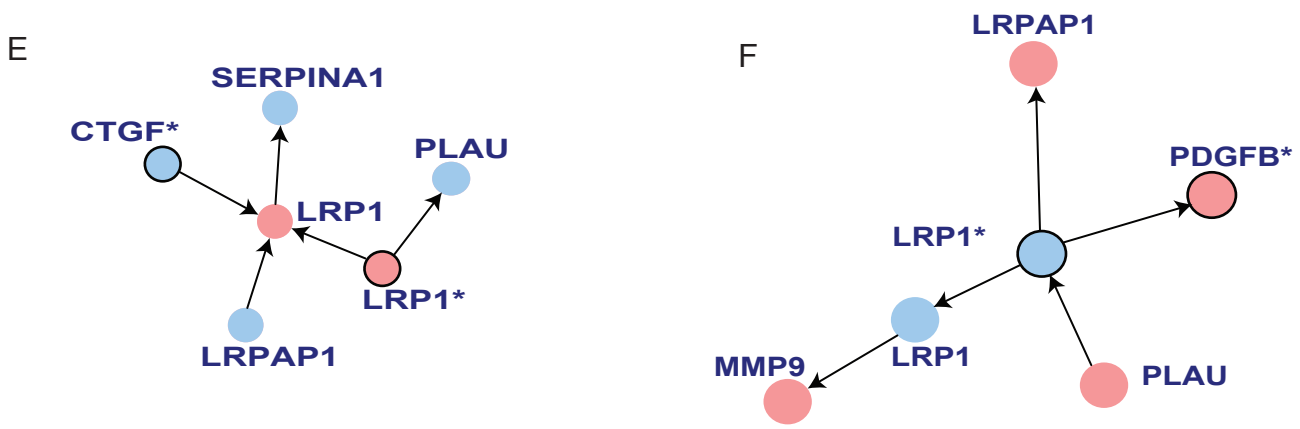
LRP1 module by LRnetST in 10x genomics data sample A4



LRP1 module by LRnetST in 10x genomics data sample A5



LRP1 module by LRnetST in 10x genomics data sample A12



Tumor spot

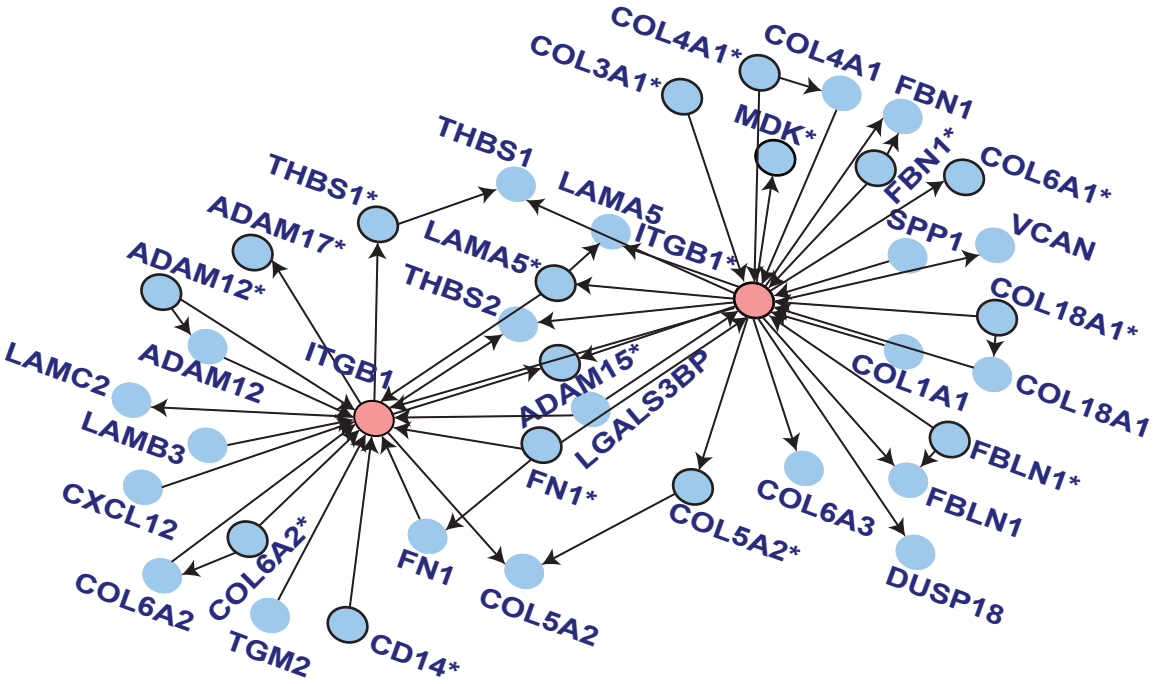
- Continuous component of the genes.
- Binary component of the genes.

Neighbor spot

- Continuous component of the genes.
- Binary component of the genes.

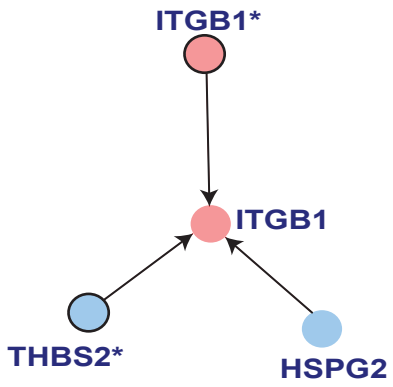
A

ITGB1 module by LRnetST
in 10x genomics data sample A4



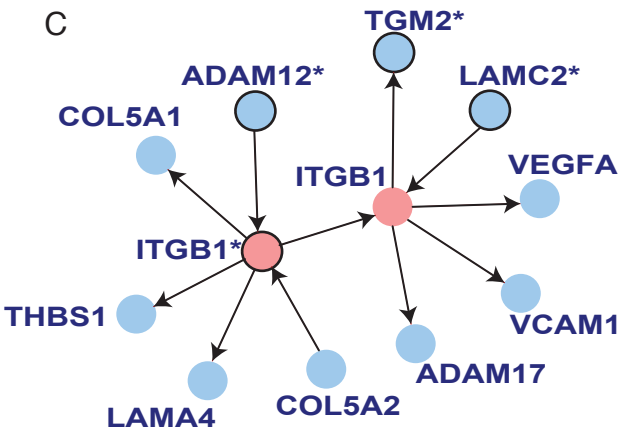
B

ITGB1 module by LRnetST
in 10x genomics data sample A5

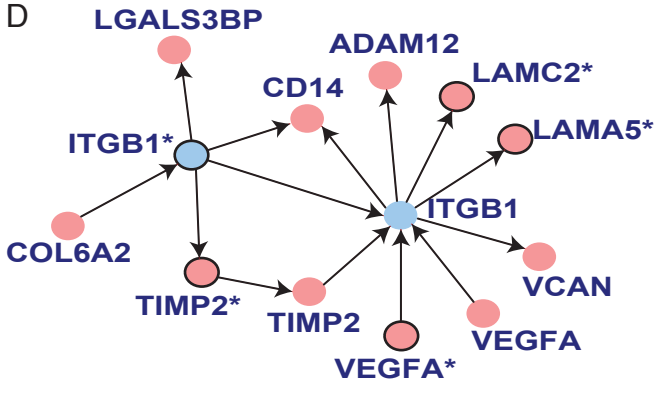


ITGB1 module by LRnetST
in 10x genomics data sample A12

C



D

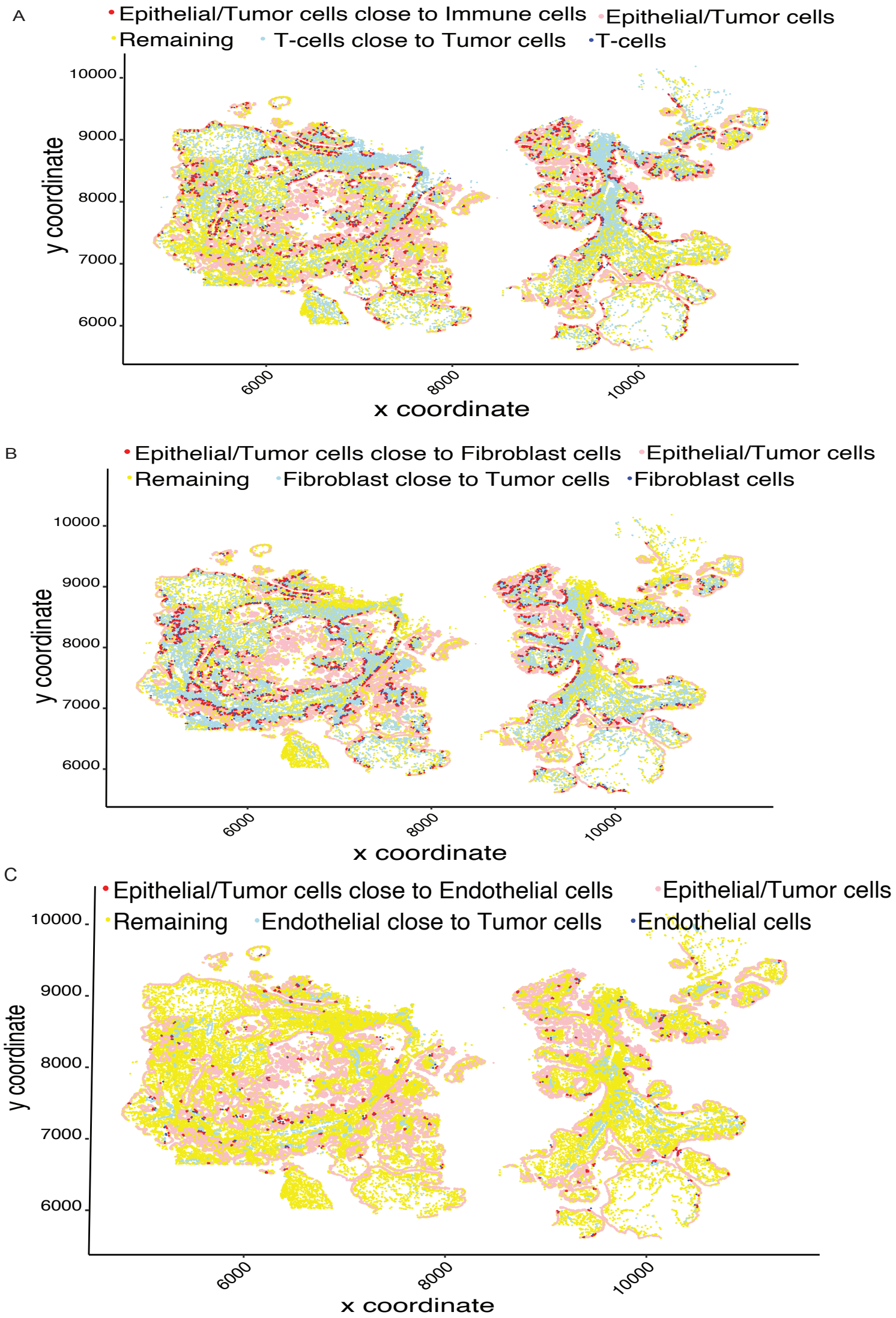


Tumor spot

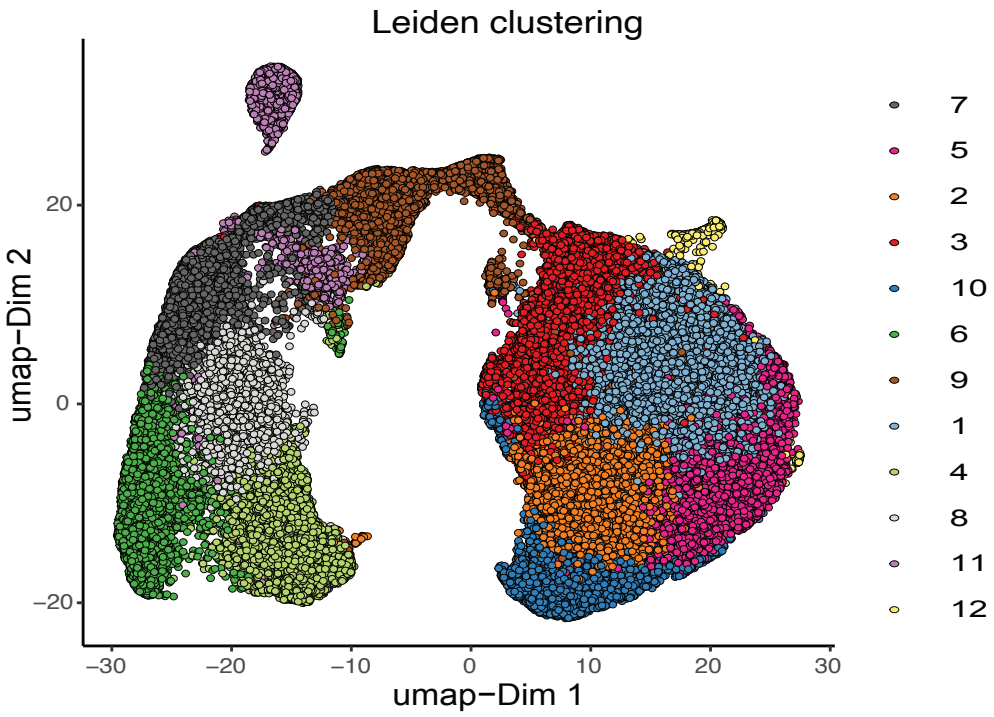
- Continuous component of the genes.
- Binary component of the genes.

Neighbor spot

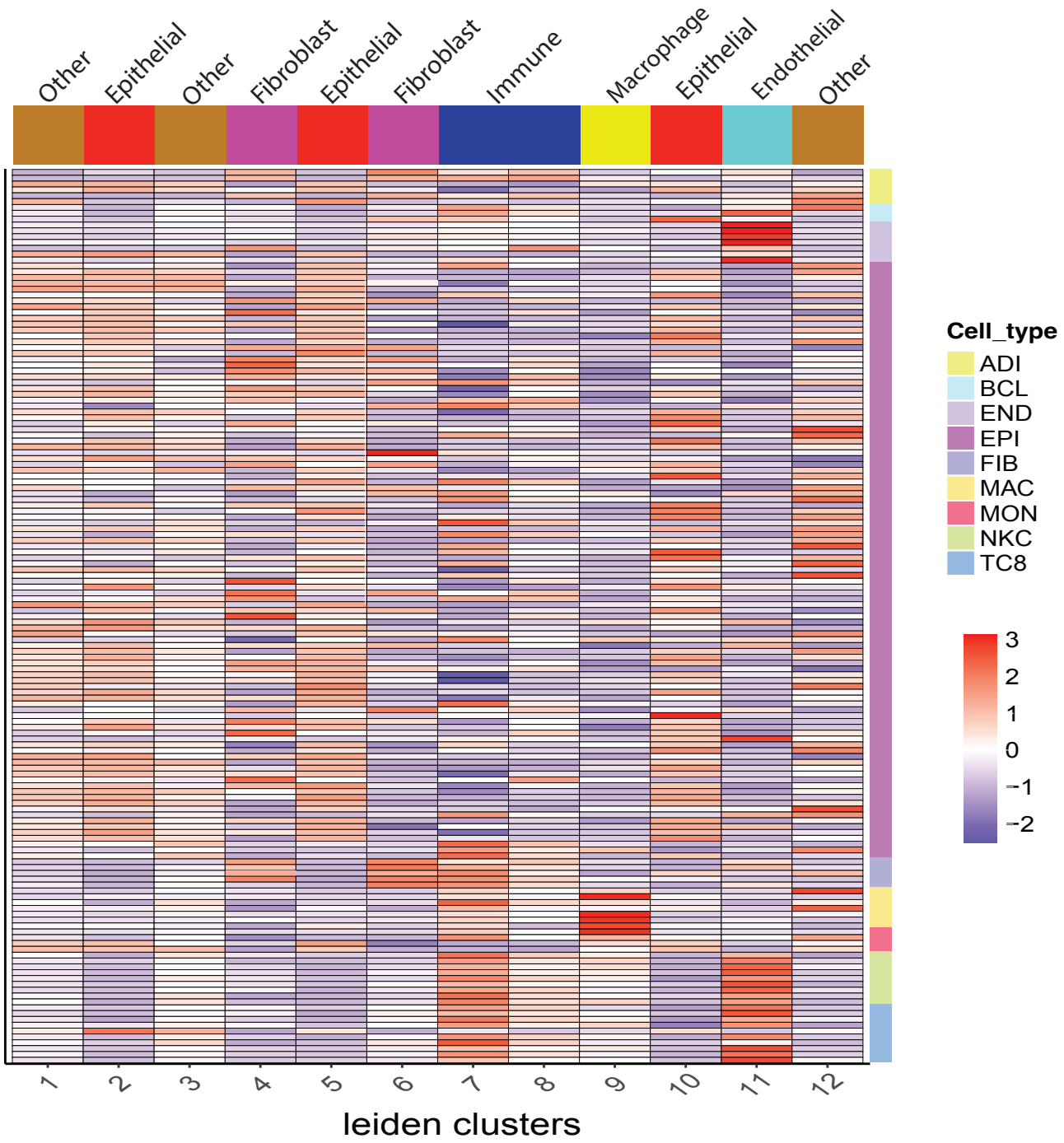
- Continuous component of the genes.
- Binary component of the genes.


A
Epithelial/Tumor cells close to Immune cells
Epithelial/Tumor cells
Remaining
T-cells close to Tumor cells
T-cells
y coordinate
x coordinate
B
Epithelial/Tumor cells close to Fibroblast cells
Epithelial/Tumor cells
Remaining
Fibroblast close to Tumor cells
Fibroblast cells
y coordinate
x coordinate
C
Epithelial/Tumor cells close to Endothelial cells
Epithelial/Tumor cells
Remaining
Endothelial close to Tumor cells
Endothelial cells
y coordinate
x coordinate

(A)


Leiden clustering
umap−Dim 2
umap−Dim 1
20
0
−20
−30
−20
−10
0
10
20
30
7
5
2
3
10
6
9
1
4
8
11
12

(B)


Other
Epithelial
Other
Fibroblast
Epithelial
Fibroblast
Immune
Macrophage
Epithelial
Endothelial
Other
1
2
3
4
5
6
7
8
9
10
11
12
leiden clusters
Cell_type
ADI
BCL
END
EPI
FIB
MAC
MON
NKC
TC8
3
2
1
0
−1
−2