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

Shrabanti Chowdhury[1*], Sammy Ferri-Borgogno[2*], Peng Yang[3,5], Wenyi Wang[3], Jie Peng[4], Samuel Mok[2†], Pei Wang[1†]

[1]*Department of Genetics and Genomic Sciences,*
Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[2]*Department of Gynecologic Oncology and Reproductive Medicine, Division of Surgery,*
The University of Texas MD Anderson Cancer Center, Houston, TX
[3]*Department of Bioinformatics and Computational Biology,*
The University of Texas MD Anderson Cancer Center, Houston, TX
[4]*Department of Statistics, University of California, Davis, CA*
[5]*Department of Statistics, Rice University, Houston, TX*

Abstract

A critical bottleneck step towards understanding the immune activation and suppression mechanisms in tumor samples is to identify transcriptional signals underlying the cell-cell communication between tumor cells and immune/stromal cells in tumor microenvironments (TME). Cell-to-cell communication extensively relies on interactions between secreted ligands and cell-surface receptors, which create a highly connected signaling network through many ligand-receptor paths. The latest advance in in situ omics analyses, such as spatial transcriptomic (ST) analysis, provide unique opportunities to directly characterize ligand-receptor signaling networks that powers cell-cell communication, which has not been feasible based on either bulk or single-cell omics data. In this paper, we focus on high grade serous ovarian cancer (HGSC), and propose a novel statistical method, `DAGBagST`, to characterize the ligand-receptor interaction networks between adjacent tumor and stroma cells in ovarian tumors based on spatial transcriptomic data. `DAGBagST` utilizes a directed acyclic graph (DAG) model with a novel approach 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. We applied `DAGBagST` to ST datasets of tumor samples from four HGSC patients, and identified common and distinct ligand-receptor regulations between adjacent tumor and stromal grids across multiple tumors. These results cast light on biological processes relating to the communication between tumor and TME cells in these ovarian tumors. An open-source R package for performing `DAGBagST` will be shared upon publication.

$*$ Co-first author

† Co-corresponding author

Keywords: Spatial transcriptomics data, ligand-receptor network, Hill climbing, Bootstrap aggregation, Prior

1 Introduction

High grade serous ovarian cancer (HGSC) is the most lethal gynecologic malignancy, and its daunting overall survival has not changed significantly for decades. Despite the tremendous efforts on reproducible delineation of HGSC genomic landscape through large-scale next-generation sequencing profilings [14, 16, 30, 37], there is still a lack of therapeutically actionable prediction biomarkers or treatment targets for HGSC patients with chemoresistance. A major challenge in fully understanding the mechanisms of tumor progression and chemoresistance of HGSC is its high intra-tumor heterogeneity, which encompasses tumor clonal heterogeneity and tissue architecture heterogeneity. The latter is reflected by the heterogeneous stromal and immune cell population in the ovarian tumor microenvironment (TME) [36]. 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 therapeutic resistance [32]. However, the molecular mechanisms that shape these cell-cell interactions in HGSC are still largely unexplored.

A major type of cell-cell signaling is powered through interactions between ligands of one cell and cognate receptors of a neighboring cell. The latest development of spatial transcriptomic (ST) profiling technology provides the 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 based on spatial transcriptomic data collected using the ST platform [26] from four HGSC ovarian samples.

The power of using network-based system learning approaches to detect disease relevant regulation patterns among genes or pathways has been demonstrated in many disease studies including cancer, COPD, and others [24, 2, 18, 10, 35, 34]. Specifically, directed acyclic graphs (DAGs) constructed based on -omics data have recently been used to infer important gene-gene interactions [17, 38, 27, 28, 6, 19, 23]. However, existing DAG methods cannot be applied directly to ST data to construct ligand-receptor networks due to multiple challenges. Firstly, one needs to incorporate both spatial information and omics profiles simultaneously in the ligand-receptor network learning, which is not readily supported by existing DAG methods/softwares. Secondly, in the sequencing output of one tissue spot in an ST experiment, often a large percentage of transcripts are missing due to both biological factors (e.g. the gene is not expressed in the cell population of that grid) and technical factors (e.g. the read depth of the experiment is limited). The resulting zero-inflated distributions of expression measurements impose statistical and computational challenges in effective network inference and mechanistic interpretation. Thirdly, UMI count levels of different spots of one tissue sample are subject to both technical variations such as different library sizes, and biological variations such as cell type compositions. Thus normalization across spots needs to be properly implemented before downstream analysis.

To address these challenges, we developed `DagBagST`, a new tool to construct ligand-receptor networks based on ST data. In the `DagBagST` pipeline, we introduce the Neighbor Integrated Matrix (NIM) to integrate the spatial information and the molecular information in the ST data. `DagBagST` then uses binary variables alongside the continuous variables for every ligand/receptor in the node space, such that (i) the binary indicator represents the on/off status of the ligands/receptors, i.e. whether active/expressed or silent/not expressed, while (ii) given the ligand/receptor is expressed, the continuous variable represents the abundance of ligand/receptor expression. Such a coding strategy not only models the zero inflation in the ST data, but

also achieves better power to detect those interactions that are largely signaled through active/inactive statuses of ligands/receptors. To account for variation in library sizes across grids in one ST experiment, `DagBagST` employs an aggregating framework that couples a bootstrap (bagging) procedure of DAG learning with the downsampling based normalization to achieve better control on false edge detection. This aggregation framework also enables `DAGBagST` to flexibly take prior information of edge directions, such as those from existing ligand-receptor databases.

We apply the proposed `DagBagST` pipeline to a ST data set of 4 HGSC samples. We make use of the known ligand-receptor regulation information from relevant databases as prior information to constrain the searching space when constructing DAGs. The inferred ligand-receptor networks reveal a common hub node, LRP1: the expression levels of LRP1 in grids enriched with tumor cells were (significantly) associated with the expression of PSAP, A2M, CALR and LRPAP1 in adjacent grids enriched with stroma cells in at least three out of the four tumors. These results suggest possible important immunological functions of LRP1 and its interacting ligands, in HGSC.

The rest of the paper is organized as follows. In Section 2 we provide a detailed description of the proposed `DAGBagST` method. In Section 3 we present the data generation and preprocessing details. Section 4 presents the application of `DAGBagST` for constructing ligand-recetor network based on ST data from four HGSC samples. We conclude the paper with some discussion in Section 5.

2 Method

We first provide a brief review of Directed Acyclic Graph models, and then introduce the new method `DagBagST`.

2.1 Background of Directed Acyclic Graph

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. In a DAG model, the node set corresponds to a set of random variables and the edge set encodes the conditional dependence relationships among these random variables. DAG structure learning amounts to identifying the parent(s) 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 ([33]). 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 $x_1 \rightarrow x_3 \in \mathbb{E}, x_2 \rightarrow x_3 \in \mathbb{E}$, 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., [8]), *constraint-based* methods , e.g., PC algorithm (`PC-alg`) [33, 25, 11] and *hybrid* methods, e.g., Max-Min Hill Climbing (`MMHC`) [31].

2.2 `DAGBagST`

In this section we present a new tool - `DAGBagST` - for learning ligand-receptor interaction networks based on spatially resolved transcriptomic data. As mentioned earlier, this task involves a few new challenges. Firstly, one needs to integrate spatial information and molecular profiles in the ST data in order to characterize interactions between ligands and receptors from the neighboring grids (spots). Secondly, ST data has high dropout (missing) rates due to failure of the capture or amplification, which then poses statistical and computational challenges in the downstream network analysis. Moreover, the UMI count levels often vary

a lot across different ST spots. Thus one needs to perform normalization across the ST spots properly to account for the different sources of variations.

To address these challenges, DAGBagST first creates Neighbor Integrated Matrices to integrate the spatial information and the molecular information in the ST data. It then codes each gene using a binary variable together with a continuous variable, such that the binary variable represents the active/inactive statuses of genes and the continuous variable represents the gene expression/abundance, given the gene is active/expressed. DAGBagST also incorporates an aggregation framework that conveniently couples the downsampling based normalization with a bootstap based network inference procedure. Figure 1 illustrates the flow of the DAGBagST pipeline. And below we elaborate the details of these steps.

2.2.1 Introduce Neighbor Integrated Matrices

In the DAGBagST pipeline, we construct a *Neighbour Integrated Matrix* (NIM) from the given gene expression data integrating the spatial and the molecular information. Specifically, for a given tumor sample, grid points in the ST data are first classified into two classes: enriched or not enriched of tumor cells. We then identify tumor cell enriched grid points that sit on the boundary of the tumor regions. We refer to such spots as the index spots and refer to their closest non-enriched spots as neighbor spots. Finally, we derive the NIM from the ST data by stacking the original gene expression matrix of the index spots with the gene expression matrix from the neighbor spots. The resulting NIM has an expanded feature space where the expression level/active status of a gene from the index spot and its neighbor spot, respectively, are treated as two features (Figure 1). The rationale behind the NIM is that we are interested in characterizing whether/how receptors in non-tumor cells (neighbor spots) in the micro-environment are affected by the ligands from the neighboring tumor cells (index spots) or vice versa.

2.2.2 Account for zero inflated distributions

To account for the high missing rates in the ST spots, DAGBagST uses two nodes, one for a binary variable and another for a continuous variable, to represent each gene in the network. The likelihood of the continuous variables are evaluated only when the corresponding binary variables take the value 1 (meaning active). This makes the binary variables natural parents of the corresponding continuous variables. Such a modeling strategy not only deals with the zero inflation in the data, but also achieves better power in detecting those interactions that are largely signaled through the active/inactive statuses of the genes.

In one of our recent work, we implemented DAGBagM [1], for DAG learning when the node space contains both continuous and binary variables. DAGBagM adopts a score-based approach where we used separate distributions for continuous and binary nodes, and thus allows the binary nodes to be either parent or child of the continuous nodes. In score-based methods, DAG space tends to grow super-exponentially with the number of nodes, making the exhaustive search of optimal models nearly infeasible. To tackle this challenge, in DAGBagM, we implemented a very efficient implementation of the hill climbing search algorithm that uses information from previous steps to facilitate the score calculation and acyclic status check in the current step and thus enables learning graphs with a large number of nodes. In DAGBagM, the score-based approach is also coupled with bootstrap aggregation to help reduce false positives by searching for stable structures.

In DAGBagST, we modified the implementation of the hill climbing algorithm in DAGBagM to optimize the new score functions. As mentioned earlier, the binary variables serves as natural parents of the corresponding continuous variables of the same genes. Specifically, for a continuous node, denoted by X, at each search step, its score is calculated by regressing X onto its current parent(s), using only the non-zero data points, i.e., using only the samples where X is non-zero.. Specifically, for a given graph $\mathcal{G}$,

$score_{BIC_X} := n_X \log(RSS_X^G/n_X) + |pa_X^{\mathcal{G}}| \log(n_X)$, where, RSS_X^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 through logistic regression:

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

where $\hat{p}_k = P\left(Y_k = 1|pa_{Y,k}^{\mathcal{G}}\right) = \frac{\exp(\hat{\gamma}_0+\hat{\boldsymbol{\gamma}}^T pa_{Y,k}^{\mathcal{G}})}{1+\exp(\hat{\gamma}_0+\hat{\boldsymbol{\gamma}}^T pa_{Y,k}^{\mathcal{G}})}$, $\hat{\gamma}_0$ 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: $score_{BIC}(\mathcal{G} : \mathbb{D}) = \sum_{X:\text{continuous nodes}} score_{BIC_X} + \sum_{Y:\text{binary nodes}} score_{BIC_Y}$, where $\mathbb{D}$ denotes the data.

2.2.3 Downsampling Based Normalization and Boostrap Based Model Aggregation

To account for the varying library sizes or cell type compositions across different tissue spots, `DagBagST` employs an aggregating framework that couples a bootstrap (bagging) procedure of DAG learning with the downsampling based normalization to achieve better control on false edge detection. Specifically we apply downsampling normalization on the original gene expression matrix, fixing the total UMI at the median level of the UMI counts of the tissue spots. We generate 100 downsampled gene expression matrices and construct NIM from each of them and then learn DAG after introducing binary variable for each continuous variable in the node space. The DAGs are aggregated following the bootstrap aggregation procedure implemented in `DAGBagM`, where we use *structural Hamming distance (SHD)* defined as the minimum number of basic operations, namely, addition, deletion and (possibly) reversal that are needed to convert one graph to another, as the metric on the DAG space. This leads to the following distance measure between two DAGs: $d(\mathcal{G}, \tilde{\mathcal{G}}) = \sum_{1\le i<j\le p} \max\{|\mathbb{A}(i,j) - \tilde{\mathbb{A}}(i,j)|, |\mathbb{A}(j,i) - \tilde{\mathbb{A}}(j,i)|\}$, where $\mathbb{A}$ and $\tilde{\mathbb{A}}$ denote the adjacency matrices of the DAGs $\mathcal{G}$ and $\tilde{\mathcal{G}}$, respectively. The adjacency matrix of a DAG is a 0-1 element matrix where the (i,j)-th element is one if there is a directed edge from the ith node to the jth node; otherwise it is zero. The *aggregation score* between a DAG $\mathcal{G}$ and an ensemble of DAGs $\mathbb{G}^e = \{\mathcal{G}_1, \cdots, \mathcal{G}_B\}$ is the average distance between $\mathcal{G}$ and the DAGs in the ensemble: $score_d(\mathcal{G} : \mathbb{G}^e) = \frac{1}{B}\sum_{b=1}^{B} d(\mathcal{G}, \mathcal{G}_b)$.. Given an ensemble (e.g., learned from the 100 down sampled data as described earlier), one can search for the DAG that minimizes the aggregation score while maintaining acyclicity by applying the hill climbing algorithm. For further details please refer to [1].

This aggregation framework also enables `DAGBagST` to flexibly take into prior information of interactions, such as those from existing ligand-receptor databases. Specifically for learning DAG from each downsampled data, we only constrain the searching space to the the prior edges among the ligands and receptors that are documented in the databases and suppress all other possible edges.

Major steps of `DAGBagST` are summarized in Figure 1 and in Algorithm 1.

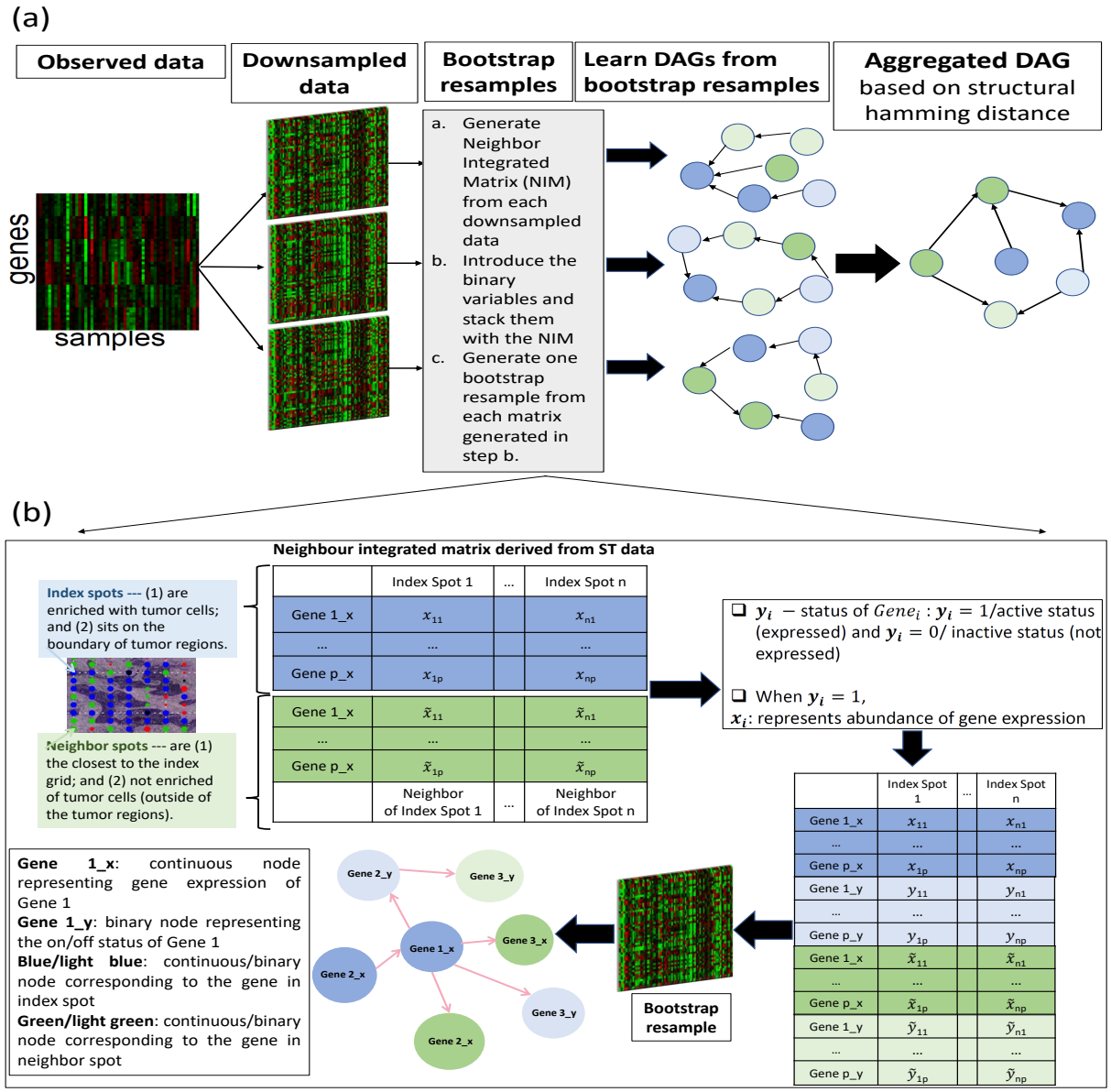

Figure 1: Overview of DAGBagST. (a) **DAGBagST pipeline**: Given the gene expression matrix B downsampled data are generated. From each downsampled data a neighbor integrated matrix is constructed and bootstrap resample is generated. DAGs are learned from each bootstrap resample and are aggregated using structural hamming distance. (b) **Construction of Neighbor Integrated matrix (NIM)**: Derive NIM based on ST data to construct Ligand-Receptor network. We first identify grids that are enriched with tumor cells and sitting on the boundary of the tumor regions, and denote these grids as Index Spots. Then for each Index Spot, we identify the closest neighboring grid(s), which are not enriched with tumor cells, referred to as the Neighbor Spot. Stacking the gene expression profiles of the index and neighbor spots results in the Neighbor Integrated Matrix.

1 `DAGBagST`: Directed acyclic graph learning through bootstrap aggregation for ST data.

Generate downsampled data: Taking the maximum number of UMI to sample as the median UMI count of the tissue spots of the original gene expression data, we apply downsampling normalization and generate B downsampled data D_b $\{b = 1, \cdots, B\}$.

Generate Neighbor Integrated Matrix: For each downsampled data D_b generate a neighbor integrated matrix M_b as follows:

- Getting the neighbor spots for the index spots: For each index spot obtain the neighbor spot being closest in terms of Euclidean distance. Filter away the index spots for which the closest neighbor spot is more than a certain distance away.
- Apply 50 % missing rate filtering on the genes and take the overlap with the ligand-receptor database – obtain p genes.
- Stack the gene expression profile of the p genes in the index spots with those from the neighbor spots. Thus for each D_b we obtain a NIM of order $2p \times n$.
- Finally introduce $2p$ binary variables for each gene in the node space, taking 1 or 0 depending on whether the gene expression is non-zero or zero. This makes the dimension of the neighbor integrated matrix $4p \times n$.

Learn ensemble: For each NIM M_b, learn a DAG $\mathcal{G}_b$ as follows:

- Initial step: Start with an empty graph $\mathcal{G}^0$: For $i = 1, \ldots, p$ calculate the score $S(\mathcal{G}^0, D_b)$ depending on the type of the node (binary or continuous). The set of possible operations (O^0) includes additions of all possible edges.
- Updating step: In the sth step ($s \geq 1$), denote the current graph by $\mathcal{G}^s$, the current score by $S(\mathcal{G}^s, D_b)$ and the current set of eligible operations by O^s (which includes any addition, deletion or reversal that does not violate acyclicity). For every eligible operation $O \in O^s$, obtain (either calculate or from the previous step) the score change: $\delta^s = S(O(\mathcal{G}^s), D_b) - S(\mathcal{G}^s, D_b)$, where $O(\mathcal{G}^s)$ denotes the graph after applying the operation O on $\mathcal{G}^s$.
- Updating/Stopping rule: Calculate the minimum score change $\delta^s_{min} = min_{O \in O^s} \delta^s$ and the corresponding operation O. If $\delta^s_{min} \geq 0$, stop and take $\mathcal{G}_b = \mathcal{G}^s$ as the estimated graph on b^{th} downsampled data; else set $\mathcal{G}^{s+1} = O(\mathcal{G}^s)$ and $s \leftarrow s + 1$ and repeat the updating step.

Aggregation: $\mathcal{G}^* = \text{argmin}_{\mathcal{G} \in \mathbb{G}(\mathbb{V})} score_d(\mathcal{G} : \mathbb{G}^e)$, where $\mathbb{G}(\mathbb{V})$ is the DAG space with the node set $\mathbb{V}$, and $\mathbb{G}^e = \{\mathcal{G}_b,\ b = 1, \cdots, B\}$, $score_d(\mathcal{G} : \mathbb{G}^e) = \frac{1}{B} \sum_{b=1}^{B} d(\mathcal{G}, \mathcal{G}_b)$, and $d(\cdot, \cdot)$ is a distance metric on $\mathbb{G}(\mathbb{V})$.

3 Data Generation and Preprocessing

Sample description and ST experiment: Four fresh frozen treatment-naive advanced stage HGSC samples from the ovarian tumor repository in the Department of Gynecologic Oncology and Reproductive Medicine at MDACC were used for the study. Two of the four samples were from two chemosensitive patients with long progression free survival time ($>$ 6 months); and the remaining two were from chemoresistant patients with short progression free survival time ($<$ 6 months). Spatial Transcriptomics profiling was performed using the ST platform as per manufacturer's instructions. Specifically, we utilized 6.9 x 6.5 mm square pre-spotted arrays (areas with active spots are 6.5x6.1mm, Fig 2a and Fig 3a) with 1007 specialized mRNA-capturing probes with a unique spatial barcode on the surface of glass slides, which are subsequently coated with frozen

tissues. Each spot (grid) is 100 microns in diameter, and the distance from the center of one spot to the center of another is approximately 200 microns, allowing a spatial resolution of 10-40 cells/grid (depending on the cell type). Tissue permeabilization, removal and cDNA synthesis was then performed directly on the tissue section. The pooled barcode specific spatially transcribed cDNA libraries for each sample were sequenced using Illumina NextSeq500 flow cells at the Advanced Technology Genomics Core at MD Anderson.

ST data preprocessing and summary: RNAseq data from cDNA libraries derived from each probe is "de-multiplexed" such that all individual sequencing reads are assigned to their spots of origin. Specifically, the output UMI (unique molecular identifier) count matrix for each sample is obtained using ST pipeline (version 1.80.1) following the recommended default parameter setting1 from FASTQ files [13]. Human genome reference (GRCh38) was used for read alignment and corresponding genome index was generated using STAR2 (version 2.7.9a) [5]. The preprocessed ST data contains expression profiles of 930-1007 grids per sample. The number of genes with non-zero UMI per grid ranges 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.

4 Result

We applied `DAGbagST` to each ST slide separately to construct tumor-specific ligand-receptor networks. Firstly, based on tissues and cells morphology in high resolution HE images, ST grids (spots) with enriched tumor cells or stroma cells were manually annotated (see Fig. 2a, 3a, and Supplementary Figure 1). We then focused on the spots (grids) enriched of tumor cells (Tumor and Tumor/Stroma in Fig 2a and 3a), and identified their neighboring spots (grids) enriched of stromal cells within 2-grid-units to construct the Neighbor Integrative Matrices in `DAGBagST` (see Fig 2b and 3b). We only considered the genes that were documented in the ligand-receptor interaction database [22], and further filtered out those detected in less than 50% of the spots (grids) in one sample. The median of the total UMIs per spot grid in each sample was used for the downsampling normalization in `DAGBagST`. Table 1 gives the detailed numbers of the tumor and stroma spots (grids), as well as the numbers of adjacent tumor-stromal spot (grid) 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 `DAGBagST`, we built four tumor specific ligand-receptor networks, one for each of the four HGSC samples. As described in the Method, `DAGBagST` introduces binary variables for each gene to indicate the on/off status of the ligands/receptors, which helps to account for zero inflation in the ST data. In addition, `DAGBagST` is flexible in taking into prior information on edges, which enables us to constrain the search space to the ligand-receptor interaction edges documented in the ligand-receptor interaction database [22] for effective dimension reduction.

Table 1: Numbers of spots enriched of tumor and/or stromal cells, and number of ligand/receptor genes in ST data sets of four tumors.

Tumor sample ID	PFS	# Spots (Tumor and Tumor/Stromal)	# Spots labelled as Stromal	# Adjacent tumor-stromal spot pairs (sample size)	# Ligand and receptor genes
4	$>$ 6 months	240	350	202	184
5	$>$ 6 months	51	147	51	111
10	$<$ 6 months	424	183	238	230
12	$<$ 6 months	359	118	236	196

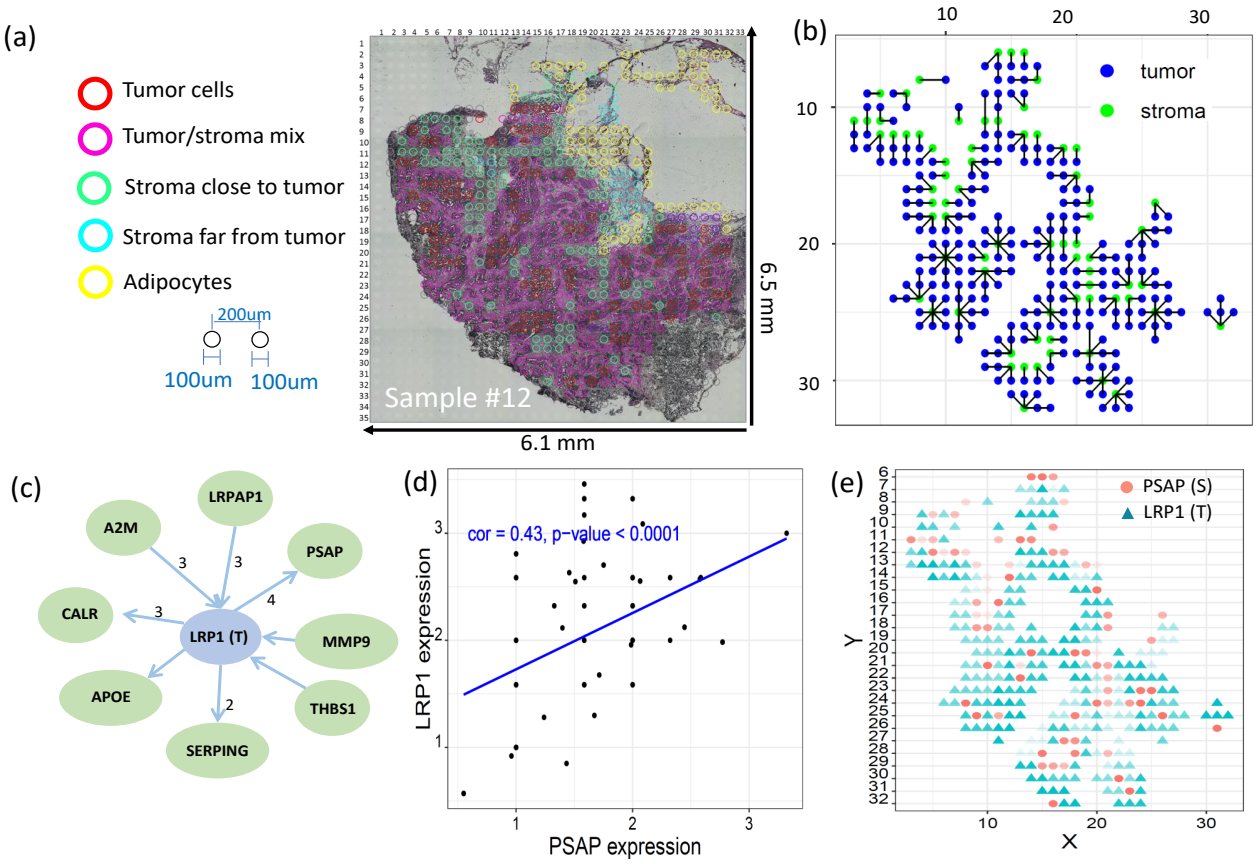

Figure 2: (a) HE image for sample #12 and the selected grids (spots) on the tissue slide for the ST profiling. Enriched cell types in each spot grid were manually annotated. (b) Pairs of adjacent spots grids that were enriched of tumor and stromal cells, respectively. These adjacent pairs of spots grids were used to construct the Neighbor Integrative Matrices in `DAGBagST`. (c) Neighbor nodes of LRP1 in the inferred ligand-receptor network for sample 12. If an edge is detected in the Ligand-receptor networks of different tumors, the numbers of the tumors (networks) sharing the edge are annotated on the lines. (d) Scatter plot showing the association between LRP1 expression levels in Tumor spots grids and the PSAP expression levels in the matching neighbor stromal spots grids as indicated in panel (b). (e) Spatial distribution of expression levels of LRP1 and PSAP. LRP1 levels (green triangular) were illustrated only for the Tumor spots grids, while PSAP levels (red triangular) were illustrated only for the adjacent stroma spots. The deeper the color is, the higher the expression levels are.

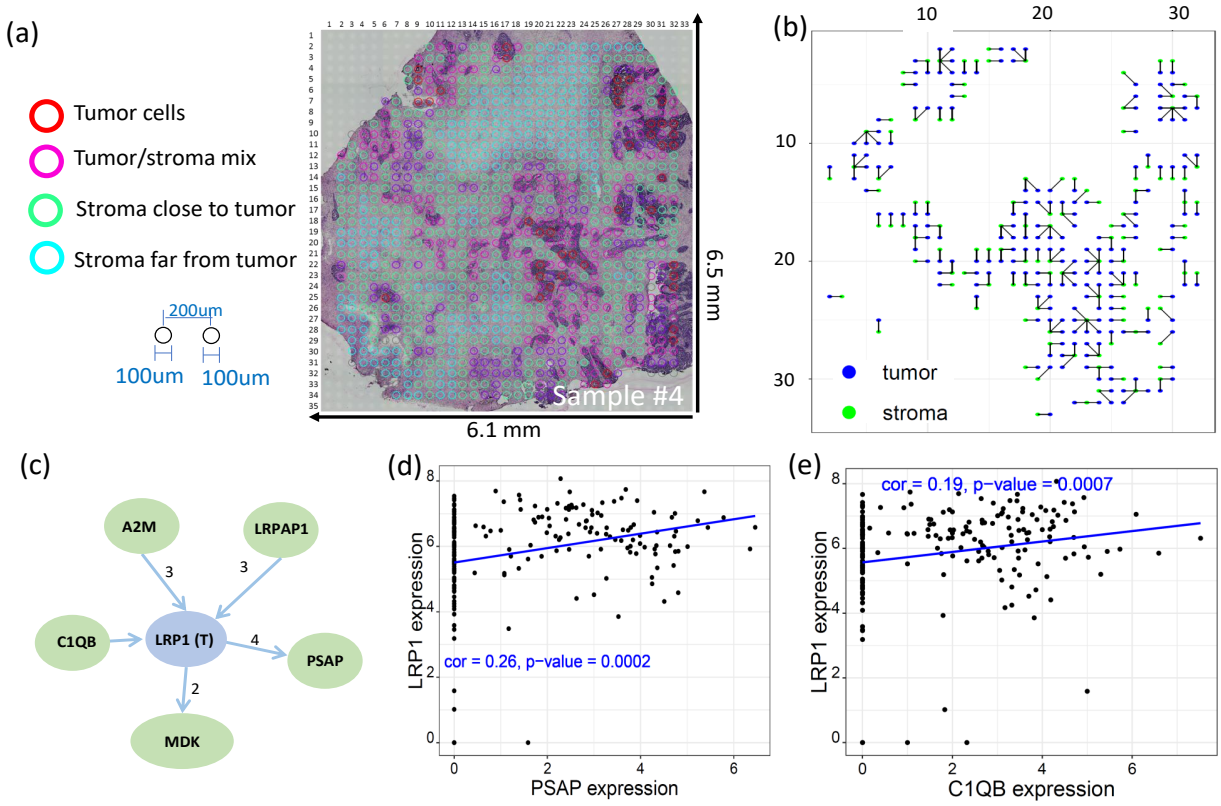

Figure 3: (a) HE image for sample #4 and the selected spots on the tissue slide for the ST profiling. Enriched cell types in each spot were manually annotated. (b) Pairs of adjacent spots that were enriched of tumor and stromal cells, respectively. These adjacent pairs of spots were used to construct the Neighbor Integrative Matrices in `DAGBagST`. (c) Neighbor nodes of LRP1 in the inferred ligand-receptor network for sample 4. If an edge is detected in the Ligand-receptor networks of different tumors, the numbers of the tumors (networks) sharing the edge are annotated on the lines. (d) Scatter plot showing the association between LRP1 expression levels in Tumor spots and the PSAP expression levels in the matching neighbor stromal spots as indicated in panel (b). (e) Scatter plot showing the association between LRP1 expression levels in Tumor spots and the C1QB expression levels in the matching neighbor stromal spots as indicated in panel

The resulting ligand-receptor networks for sample 4, 5, 10 and 12 contained 55, 54, 134, and 102 ligand-receptor regulation edges respectively (Supplementary Figure 2). Both common and distinctive interaction patterns were observed across the four tumor samples. Specifically, LRP1 appeared to be a hub node in all four networks, having a degree of 5, 7, 4 and 8, for sample 4, 5, 10 and 12, respectively. The interaction between the $\text{LRP1}_{\text{Tumor}}$ (LRP1 levels in Tumor spots grids) and the $\text{PSAP}_{\text{Stroma}}$ (PSAP levels in Stroma spots grids) was the only edge detected in all four tumor samples (Fig 2c-2e, and Fig 3c-3d). Besides PSAP, multiple other interactions of $\text{LRP1}_{\text{Tumor}}$ were also shared across tumors. For example, $\text{LRP1}_{\text{Tumor}}$ is inferred to interact with A2M, CALR and LRPAP1 in the neighboring Stroma spots grids in three out of the four tumors. On the other hand, there are also tumor specific interactions such as $\text{LRP1}_{\text{Tumor}}$ —$\text{C1QB}_{\text{Stroma}}$ in sample 4 (Fig 3c, 3e) and $\text{LRP1}_{\text{Tumor}}$ — $\text{APOE}_{\text{Stroma}}$ in sample 12 (Fig 2c).

5 Discussion

To gain insights on the molecular basis underlying cell-cell interactions in tumor microenvironment, in this paper, we introduced a new statistical method — `DAGBagST` to infer ligand-receptor interaction networks among adjacent cells based on spatial transcriptomic data. Built upon the latest development of directed acyclic graph bayesian graphical models, `DAGBagST` employs novel treatments to properly handle the spatial structure and high dropout rates of ST data. It also leverages existing ligand-receptor regulation databases as prior information, and uses a bootstrap aggregation strategy to achieve robust network edge estimation.

We applied `DAGBagST` to an ST data set of four high grade serous ovarian cancer samples and constructed patient specific ligand-receptor networks. The results reveal both common and distinct ligand-receptorion regulations between adjacent grid spots grids enriched with tumor and stromal cells separately across different patients.

LRP1 has been identified as a hub node in all four ligand-receptor networks. LRP1 (low-density lipoprotein receptor related protein 1) was initially identified as a tumor suppressor, as it exerts a dramatic control of tumor cell plasticity and migratory capacities, It is believed to involve in the control of the invasive potential of breast cancer tumor cells [4]. Interestingly, interaction between $\text{LRP1}_{\text{Tumor}}$ and $\text{PSAP}_{\text{Stroma}}$ was detected in all four samples. pSAP (prosaposin) is the common precursor to the four saposins A-D [7, 15], essential players for sphingolipid degradation and membrane digestion [3]. PSAP is intracellularly targeted to lysosomes [12], and it can also be secreted and re-endocytosed by LRP1 [9]. The single- cell RNAseq study by Cell Type Atlas at the Human Protein Atlas [29, 21] suggests significantly higher expression of PSAP in macrophages (including Hofbauer, a type of macrophage in placenta, and Kupffer, a type of liver macrophages) and monocyte cells than in other cell types. Saposins and its precursor pSAP have been suggested to play a significant role in lipid antigen presentation, a mechanism to activate lipid-specific T cells [3]. For example, human pSAP-deficient fibroblasts expressing CD1b failed to present mycolic acid, GMM, and LAM for activation of antigen-specific CD1b-restricted T cell clones [20]. It was also hypothesized that pSAP-dependent processing of apoptosis vesicles contribute to antigen delivery in DCs and subsequent CD8+ T cell responses [20]. Our finding further suggests possible important immunological functions of pSAP, as well as LRP1, in HGSC.

Another interesting ligand inferred to interact with LRP1 is C1QB. The interaction between $\text{LRP1}_{\text{Tumor}}$ and $\text{C1QB}_{\text{Stroma}}$ was, however, only detected in tumor $\#4$ from a chemosensitive patient (PFS $>$ 6 months). C1QB is a pattern recognition molecule of innate immunity and can be locally synthesized by macrophages and dendritic cells (cite). The connection between LRP1 and C1QB detected here suggests a potential role of LRP1-C1QB crosstalk network in modulating immune response which may lead to improved survival in HGSC patients.

In summary, spatial transcriptomic data provide an unprecedented opportunity to pull back the curtain of the molecular mechanism underlying the cell-cell interaction in the tumor microenvironment. The proposed `DAGBagST` method is a useful tool to construct the global ligand-receptor interaction networks based on ST data. Results from this type of analyses help to reveal important players driving cell-cell interactions, and potentially can lead to identification of new predictive biomarkers or therapeutic targets.

6 Acknowledgements and Funding

This work is supported by grants (U01 CA214114, U24 CA210993) from the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC).

References

[1] S. Chowdhury et al. Dagbagm: Learning directed acyclic graphs of mixed variables with an application to identify prognostic protein biomarkers in ovarian cancer. *bioRxiv*, 2020.

[2] D.J. Clark et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, 179(4):964–983, 2019.

[3] A. Darmoise, P. Maschmeyer, and F. Winau. The immunological functions of saposins. *Adv Immunol*, 105:25 – 62, 2010.

[4] S. Dedieu and B. Langlois. Lrp-1: a new modulator of cytoskeleton dynamics and adhesive complex turnover in cancer cells. *Cell Adh Migr*, 2:77–80, 2008.

[5] A. Dobin et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15 – 21, 2013.

[6] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[7] W. Fürst, W. Machleidt, and K. Sandhoff. The precursor of sulfatide activator protein is processed to three different proteins. *Biol Chem Hoppe Seyler*, 369(5):317 – 328, 1988.

[8] D. Geiger and D. Heckerman. Learning gaussian networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 235–243. Morgan Kaufmann Publishers Inc., 1994.

[9] T. Hiesberger, S. Hüttler, A. Rohlmann, W. Schneider, K. Sandhoff, and J. Herz. Cellular uptake of saposin (sap) precursor and lysosomal delivery by the low density lipoprotein receptor-related protein (lrp). *EMBO J*, 17(16):4617 – 4625, 1998.

[10] L. Jostins et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119 – 124, 2012.

[11] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

[12] S. Lefrancois, J. Zeng, A. J. Hassan, M. Canuel, and C. R. Morales. The lysosomal trafficking of sphingolipid activator proteins (saps) is mediated by sortilin. *EMBO J*, 22(24):6430 – 6437, 2003.

[13] J. F. Navarro et al. St pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics*, 33(16):2591–2593, 2017.

[14] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609 – 615, 2011.

[15] J. S. O'Brien, K. A. Kretz, N. Dewji, D. A. Wenger, F. Esch, and A. L. Fluharty. Coding of two sphingolipid activator proteins (sap-1 and sap-2) by same genetic locus. *Science*, 241(4869):1098 – 1101, 1988.

[16] A. M. Patch et al. Whole-genome characterization of chemo-resistant ovarian cancer. *Nature*, 521(7553):489 – 494, 2015.

[17] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[18] F. Petralia et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. *Cell*, 183(7):1962–1985, 2020.

[19] D. PeÕer, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1):S215–S224, 2001.

[20] F. Pontén, K. Jirström, and M. Uhlen. Saposin c is required for lipid presentation by human cd1b. *Nat Immunol*, 5(2):169 – 174, 2004.

[21] F. Pontén, K. Jirström, and M. Uhlen. The human protein atlas–a tool for pathology. *J Pathol*, 216(4):387 – 93, 2008.

[22] J. Ramilowski, T. Goldberg, J. Harshbarger, et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat Commun*, 6(7866):755 – 765, 2015.

[23] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science Signalling*, 308(5721):523, 2005.

[24] S. Sivendran et al. Dissection of immune gene networks in primary melanoma tumors critical for antitumor surveillance of patients with stage ii–iii resectable disease. *Journal of Investigative Dermatology*, 134(8):2202–2211, 2014.

[25] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2001.

[26] P. L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[27] W. H. Sung, C. Gong, C. Myun-Seok, and Z. Hua. 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.

[28] W. H. Sung and Z. Hua. Estimation of sparse directed acyclic graphs for multivariate counts data. *Biometrics*, 72(3):791–803, 2016.

[29] P. J. Thul et al. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017.

[30] R. W. Tothill et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.*, 14:5198 – 5208, 2008.

[31] I. Tsamardinos, L.E. Brown, and C.F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[32] T. Tsujikawa, J. Mitsuda, H. Ogi, A. Miyagawa-Hayashino, E. Konishi, K. Itoh, and S. Hirano. Prognostic significance of spatial immune profiles in human solid cancers. *Cancer Sci*, 111(10):3426–3434, 2020.

[33] T. Verma and J. Pearl. 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*, pages 220–227, 1991.

[34] S. Yoo et al. Integrative analysis of dna methylation and gene expression data identifies epas1 as a key regulator of copd. *PLoS Genet.*, 11(1):e1004898, 2015.

[35] B. Zhang et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707 – 720, 2013.

[36] B. Zhang et al. Revisiting ovarian cancer microenvironment: a friend or a foe? *Protein Cell*, 9(8):674–692, 2018.

[37] H. Zhang et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, 166(3):755 – 765, 2016.

[38] J. Zhu, P. Sova, Q. Xu, K. M. Dombek, E. Y. Xu, and H. et al Vu. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology*, 10(4), 2012.