Discovering paracrine regulators of cell type composition from spatial transcriptomics using SPER

Tianxiao Zhao[1,2], Adam L. Haber[2]

[1] Institute of Systems Genetics, New York University Grossman School of Medicine, New York, NY
[2] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA.

Correspondence to Adam L. Haber (ahaber@hsph.harvard.edu)

SUMMARY:
A defining characteristic of each biological tissue is its cell type composition. Many pathologies and chronic diseases are associated with perturbations from the homeostatic composition, which can lead to aberrant tissue function. Spatial transcriptomics now enables simultaneous measurement of gene expression and cell type composition, creating an opportunity to discover the transcripts most strongly associated with changes in nearby cell composition. However, no method yet exists to systematically identify such signals. Here, we develop Spatial Paired Expression Ratio (SPER), a computational approach to evaluate the spatial dependence between transcript abundance and cell type proportions in spatial transcriptomics data. We demonstrate the ability of SPER to accurately detect paracrine drivers of cellular abundance using simulated datasets. Using publicly available spatial transcriptomic data from mouse brain, we show that genes identified by SPER are highly likely to be both extracellularly secreted and to participate in known receptor-ligand interactions. This provides evidence of SPER's ability to detect compositional regulatory signals. In addition to recovering known interactions, SPER identified novel spatial associations, such as between the Wnt ligand *Rpso3* and its cognate receptor *Lgr5*, not previously known to regulate neural cell type composition in the motor cortex. SPER thus presents a general approach to discovering paracrine drivers of cellular compositional changes.

Keywords: Spatial transcriptomics, spatial data analysis, spatial compositional analysis, systems biology, computational biology, cell circuits, network biology, cellular composition, cell-cell communication.

INTRODUCTION:

Given the specialized character of each cell type, the distribution of types that compose each biological tissue is fundamental to mediating physiological functions. Accordingly, compositional changes, such as influx of immune cells, proliferation of a stem cell pool, or metaplasia of fibroblasts, are tightly regulated by paracrine signaling networks. For example, inflammation can lead to cytokine-mediated changes in cell type composition, such as the canonical 'weep and sweep' response to intestinal parasites, involving mucus-producing goblet cell hyperplasia, eosinophilia and basophilia [1]. Many diseases involve this type of change in corresponding tissues: in Alzheimer's disease for example, neuropathology is associated with aberrant proliferation of microglia and astrocytes [2], which may be mediated by paracrine CX3CL1/CX3CR1 signaling [3]; in asthma, airway remodeling driven by IL-13 leads to increased numbers of smooth muscle and goblet cells accompanied by the abnormal recruitment of immune cells [4]. Triggering factors, like environmental exposures [5] and microenvironmental signals [6], can induce cellular composition changes. To gain insight into the normal physiology of tissue composition and transformations associated with disease, it is essential to understand *which signals regulate the compositions of cell types* in a tissue.

Single-cell transcriptomics now provides the transformative ability to simultaneously measure cell-type composition and the expression of factors that regulate it. However, it is unable to determine whether the signals and the receptor cell types are sufficiently near to each other that the signaling process is possible. Spatial transcriptomics (ST) promises to remove this hurdle by simultaneously revealing (1) the expression of paracrine ligands and receptors, (2) the pattern of cell type compositions, and (3) the spatial relationships between the relevant pairs of transcripts and cell populations. ST is an overarching term for a range of methods designed to capture transcriptome-wide gene expression while retaining information about the spatial tissue context [7]. The most widely-used and relatively affordable technologies, including ST Visium [8] and Slide-seq [9], capture spatial information in voxels around 55 microns in diameter, larger than a single cell. Recovering the relative proportions of component cell types, can be straightforwardly achieved by available deconvolution algorithms. Probabilistic-based approaches, like RCTD [10], Cell2location [11], and SONAR [12], model the expression data with certain distributions under a Bayesian likelihood framework. All these methods produce data which has the constraint that in each spot all proportions sum to 1, and thus their output is compositional. When combined with gene expression data, this type of spatial compositional data can potentially be used to discover paracrine drivers of tissue transformation, but no computational tools yet exist to do so.

In lieu of appropriate methods, several studies have applied classical statistical correlations (Pearson's R and Spearman's ρ) to find such signals in spatial transcriptomics data [13-15]. However, analyzing spatial compositional data, which has a unique set of characteristics, requires specialized methods [16], since treating each individual cell type's composition as an independent measurement leads to spurious correlations [17]. In addition, there are further problems with applying classical metrics to this problem. For a given cell type A, there are two categories of transcriptional signals that show spatial dependence with its compositional distribution. The genes specifically expressed by A, or 'A-type marker genes' will show a trivial spatial relationship with the composition of A, simply because A cells are the main source of these transcripts. Far more interesting, from our point of view, is a second category of genes which show a spatial relationship with the composition of A, but are not expressed by A cells; these genes are candidate paracrine

regulators that may regulate *A*'s composition. Correlation metrics may fail to detect this second category, since the spatial signals might not exactly overlap, and indeed, may not overlap at all.

To overcome this problem, we introduce Spatial Paired Expression Ratio (SPER), a new computational approach to detect paracrine signals that regulate the composition of a target cell type, and which may therefore show a non-overlapping spatial dependence. SPER is inspired by the pair correlation function (radial distribution) in physics, which evaluates how two types of particles are spatially correlated with each other by estimating the probability of one particle appearing at a given distance to the other [18]. To denoise our spatial compositional signal while modeling its compositionality we apply co-kriging from the geostatistics field [19], prior to computing pair correlations to identify transcripts co-varying with spatial compositions. In the following sections we introduce the SPER algorithm, and then show using simulated data that it substantially outperforms all other applicable metrics – Pearson correlation, Spearman correlation, and Earth Mover's Distance (EMD). We next apply SPER to real ST profiling of adult mouse brain and show that genes identified by SPER are far more significantly enriched for both extracellular genes and known paracrine ligands, consistent with a higher probability of participating in paracrine interactions. Finally, we use the same dataset to show that SPER can identify novel, biologically significant interactions, such as the neuropeptide precerebellin (*Cbln1*) and Meis2^{+} neurons (which express its receptor *Nrxn2*), and the Wnt ligand R-spondin 3 (*Rspo3*) associated with the same Meis2^{+} neuron subset (which expresses its receptor *Lgr5*). Both novel putative paracrine signal links are identified by SPER but cannot be distinguished by other methods. Taken together, our results show that SPER offers a general approach to the detection of putative regulators of cell type composition from spatial transcriptomics data.

RESULTS:

SPER detects simulated paracrine drivers of cell type composition.

The SPER algorithm takes three inputs, normalized spatial transcriptomics matrix, cell-type spatial compositional matrix, and the spot-spot distance matrix, which can all be computed from the raw spatial transcriptomics data (**Fig. 1**). Estimation of the cell type compositional matrix can be performed using any of the available spatial deconvolution tools; we selected RCTD [10].

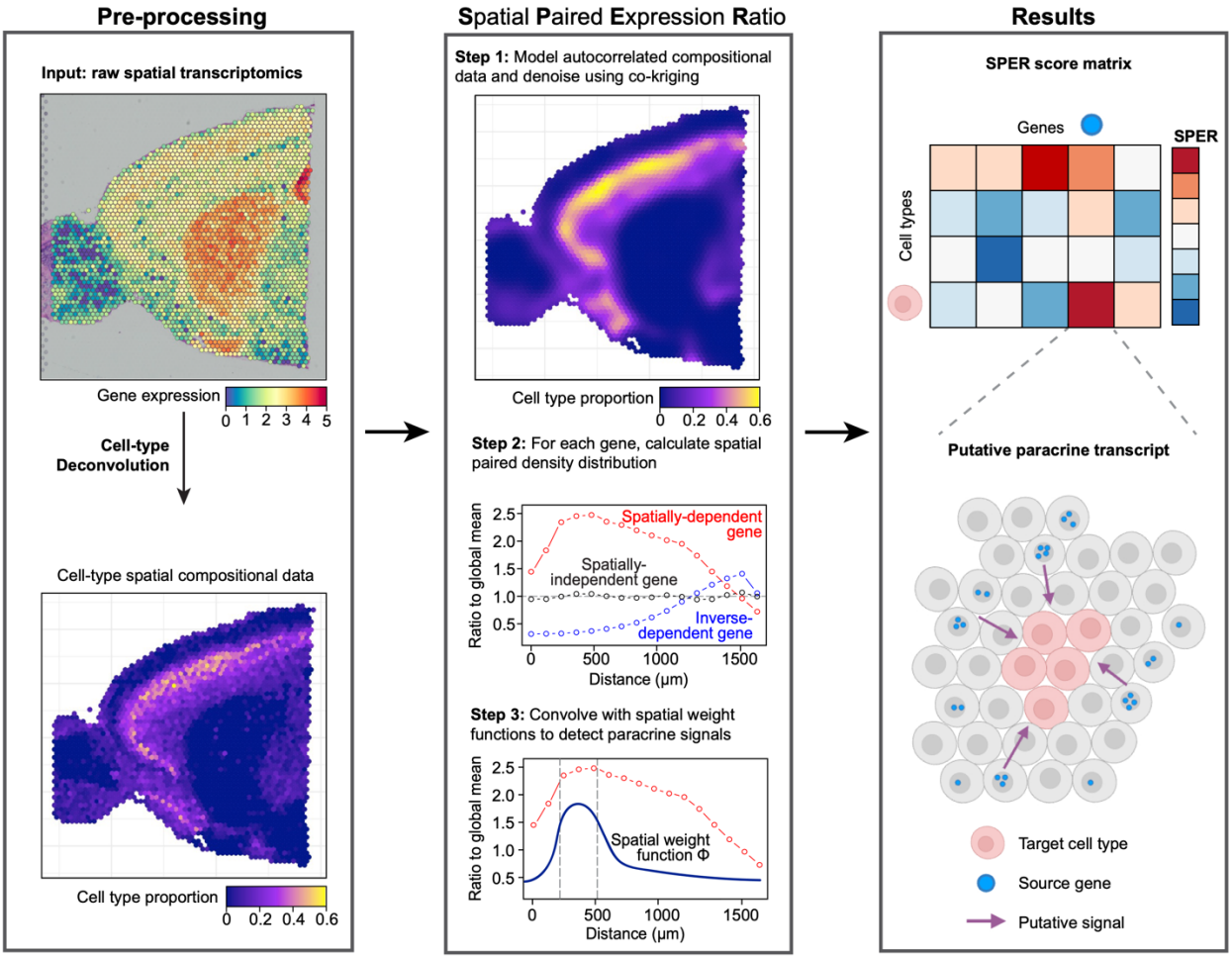


Figure 1. Schematic of spatial paired expression ratio (SPER) algorithm. The workflow of SPER: (Left) **Pre-processing**: SPER requires three different inputs: spatial transcriptomics data, spatial cell-type compositional data, and the spot-spot distance matrix. Any cell-type deconvolution method can be used to generate the spatial compositional data. (Center) **Core algorithms**: SPER models the spatial dependence between genes and their potential target cell types by measuring the expression changes over the interaction distance of a cell type-gene pair. As the first step, we apply co-kriging algorithms to model and de-noise the spatial compositional data. Next, we evaluate the spatial association by calculating a gene's ratios to its mean expression over the distance towards the target cell type. These ratios are defined as the paired expression ratios, and a spatial weight function on these ratios is applied at the last step to emphasize paracrine signal patterns. (Right) **Results**: the algorithm generates a SPER score matrix as the final output, where higher scores indicate a stronger confidence of this cell type-gene pair as a putative paracrine signal, with a schematic example transcript cell-type pair shown (bottom right).

While no method yet exists to mine ST data for paracrine signals regulating cell type composition, the major analytical field to derive inferential tools for spatial compositional data has been geostatistics. This branch of statistics focuses on modeling multivariate data that show codependency related to spatial proximity, originally developed in predicting the co-distribution of metal ores for mining applications [20]. Co-kriging is a geostatistical interpolation approach that models compositionality as well as spatial dependance between components of a mixture. It uses kernel functions to describe the pattern of variograms between each pair of components as a function of geographic lag distance [17]. SPER leverages these desirable properties of co-kriging to de-noise the spatial compositional data, modeling variograms between cell type proportions, and adjusting for spatial autocorrelation. Then, using the distance information obtained from the physical coordinates of all ST spots, calculates the paired expression ratio as a series of matrices, which includes all pairs of cell types and genes; finally, SPER convolves the ratios at each distance with a weight function to integrate over length scales of interest and obtain a scalar score representing the spatial relationship between the given cell type and the gene (**Fig. 1** and **Methods**).

To evaluate how well SPER detects paracrine regulators of cell composition, we applied it to three different simulated ST datasets (**Fig. 2**) and compared it with traditional methods of detecting spatial co-variation, specifically: Pearson correlation, Spearman correlation, and Earth Mover's Distance (EMD) [21]. To set up these synthetic benchmark datasets, we simulated 80-by-80 grids where each tile represents a spot in the physical space. Next, different cell types and gene expression patterns were assigned to tiles in the following manner. Cell type specific 'marker genes' for each cell type, only expressed in the corresponding cells, were defined across the space with the ability to 'diffuse' to its nearby neighbors. The 'paracrine signal genes', which we are interested in, can be mathematically described as genes having a similar but not directly corresponding spatial distribution as their target cell type. Such genes can be expressed either by a sole source cell type or be jointly expressed by multiple cell types. Therefore, we setup the simulations in three scenarios: simplified single-source, realistic single-source, and realistic multi-source paracrine signal patterns (**Fig. 2, left**), using radially dependent association for a simplistic and Markov random fields for more realistic simulations (**Methods**). With these simulated ST data, cell-type deconvolution methods were then applied to generate cell type compositional data. SPER, as well as other measurements, were then applied to detect the paracrine signal genes.

In the simplified single-source simulation, the signal genes are simulated to be expressed by one source cell type around the target cell type at a fixed radial distance, making their spatial distributions dependent on the cell types' but not overlapped (**Fig. 2, top-middle**). To improve the realism of the synthetic data, we next designed a more complex single-source simulation, where the cell type spatial distribution is simulated by a Markov random field [22]. Here, the target cells are set to be spatially associated but not overlapped with their corresponding source cell types in a more realistic manner (**Fig. 2, center**). Additionally, since a cell can be regulated through a signaling pathways with multiple sources, we also simulated a multi-source scenario where target cells are spatially surrounded by two source cell types expressing the signal genes (**Fig. 2, bottom-middle**).

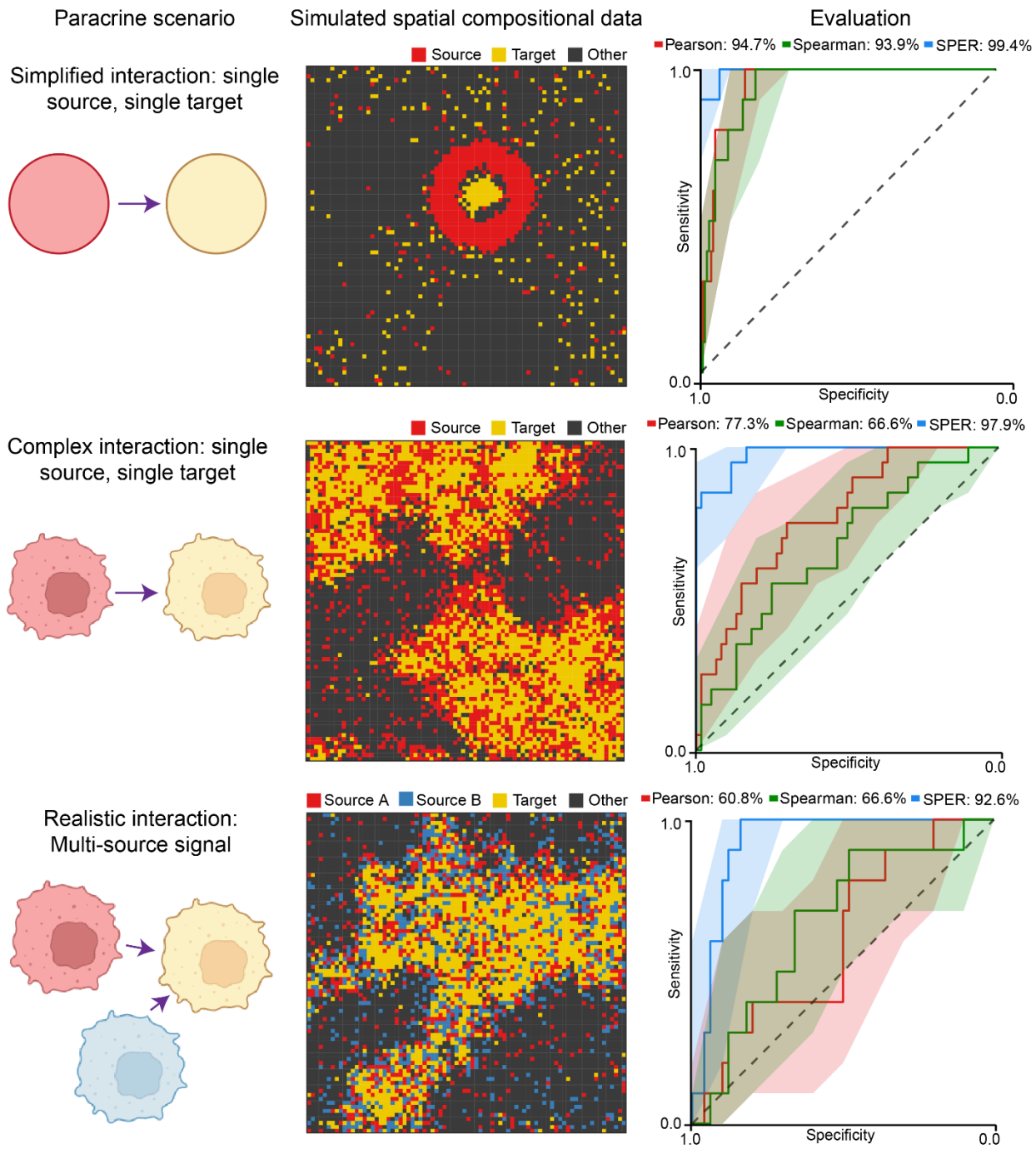


Figure 2. Evaluation of SPER on different simulated spatial transcriptomics data. **Left**. Three paracrine signal patterns - simplified single-source, realistic single-source, realistic multi-source - are simulated in this study. **Middle**. Visualizations of representative simulated spatial compositional data in an 80x80 grid, with each tile colored by its major cell type (color legend). In the realistic scenarios, Markov random fields were used to simulate the spatial dependence between the sources and target. **Right**. AUROCs are calculated to evaluate how well each metric distinguishes simulated paracrine signals from other spatially unrelated genes.

To compare the paracrine pattern to other genes, we also simulated a control set of genes that are uniformly distributed in the grid and therefore have no spatial association with any cell type. We considered the marker genes expressed by target cell types as part of control group. Thus, we can evaluate a method's performance

169 on detecting paracrine gene signals by how well it can distinguish them from the control set. Four different
170 metrics, SPER, Pearson and Spearman correlation, and Earth Mover's Distance (EMD), were applied in the
171 simulations (**Fig. 2, right**). While all other metrics completed in minutes, EMD, a method with extremely
172 high computational complexity, failed to complete within two hours. Therefore, we decide to exclude this
173 metric from the list as it is therefore impractical to apply on real data. While all three methods exhibited a
174 good detection accuracy in the simplified simulation, SPER consistently showed the best performance,
175 accurately detecting signals in all three scenarios with AUROC scores from 0.92 to 0.99. Pearson and
176 Spearman correlations, which did well in the simplified scenario, failed to capture the paracrine signals in
177 two more realistic simulations with the AUROC scores from 0.60 to 0.77. Taken together, our simulation
178 studies demonstrate the ability of SPER to detect paracrine signals that might be missed by other traditional
179 metrics, especially in a more complex and realistic settings.
180
181 **SPER outperforms classical metrics in detecting potential paracrine signals in real ST data**
182 To evaluate its performance on real data we next applied SPER, along with the other metrics, to ST profiles
183 of mouse brain generated using 10X Visium [8]. After preprocessing, 18270 unique genes were detected
184 across 2695 spots sampled from a sagittal anterior mouse brain section; cell type compositions of each spot
185 were estimated using RCTD with matched scRNA-seq as the reference [23] (**Fig. S1A**). SPER's co-kriging
186 models showed good fits to the observed variograms of top ten cell types (**Fig. S2**), and visualizing the
187 compositional data after co-kriging demonstrated its ability to remove noise (**Fig. S1B**). For each of the ten
188 neuronal cell types in the dataset, we used the scRNA-seq reference to define cell-type-specific marker
189 genes that were strongly upregulated in one particular cell type (at least two-fold higher, FDR<0.001,
190 calculated using MAST [24]). Clearly, given the source of these transcripts, we expect their expression in
191 the ST data to be spatially strongly coupled to the composition of that cell type. Consistently, using Pearson
192 and Spearman correlations, we confirmed that these marker genes showed significant spatial dependency
193 to their corresponding cell types (**Fig. S3**).
194
195 Since on the other hand, our goal is to identify paracrine signals, we hope to detect those genes that are not
196 expressed (or only lowly expressed) in the target cell type but are instead expressed by neighboring source
197 cells. We defined 'putative paracrine transcripts' as those whose spatial dependency was significantly
198 associated with a target cell type, but its expression prevalence in that same target cell type was less than
199 specified thresholds. To maximize detection of these transcripts, we adjusted the SPER ratio using the
200 scRNA-seq expression data (**Methods**). To visualize the extent to which putative paracrine transcripts were
201 enriched in each metrics top ranked genes, we plotted the metric scores against the proportion of cells
202 expressing each gene in the reference scRNA-seq data [23] for each cell type-gene pair (**Fig. 3A-B, S4**). In
203 L4 (Layer 4) neurons for example (**Fig. 3A-B**), all genes above the 95th percentile of Pearson correlation
204 also showed high expression prevalence in L4 neurons; and so, it detected no putative paracrine signals at
205 all at the 95th percentile threshold (**Fig. 3A**). In fact, most of these genes are trivially L4 marker genes. On
206 the other hand, in the case of SPER, there are no marker genes among the top 95th percentile, and the
207 majority meet our baseline criterion to be putative paracrine transcripts (**Fig. 3B**). This trend was consistent
208 across all cell types after varying the metric percentile thresholds from 75th to 95th, where SPER detected
209 markedly more putative paracrine genes whose expression prevalence is less than 20% of the target cell
210 type (**Fig. 3C**), supporting SPER's capability to detect putative paracrine regulators of cell type composition.
211

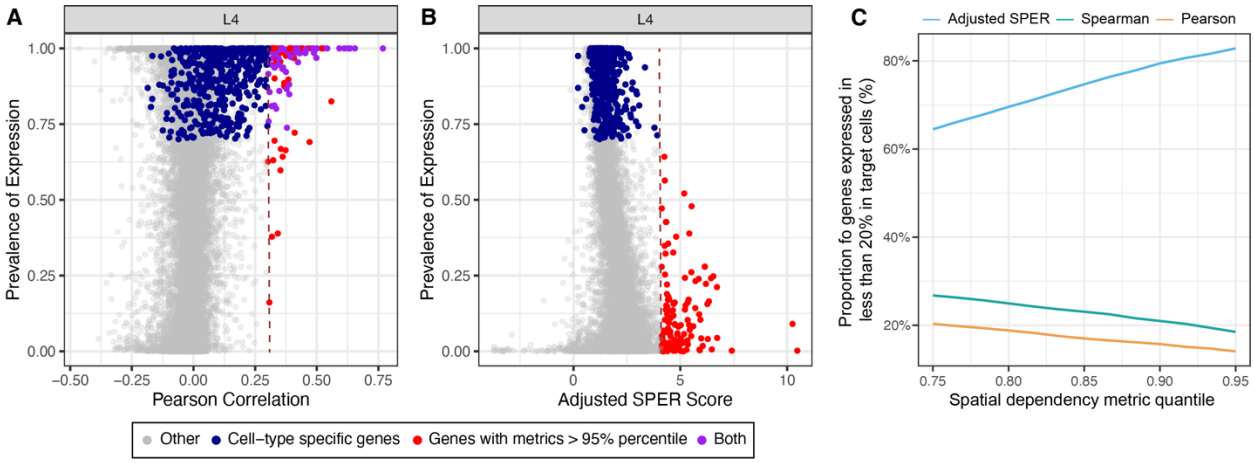


Figure 3. SPER detects non-trivial spatial relationships. (A, B) Scatter plots show the distribution of Pearson correlation (A) and SPER (B) scores with the proportion of L4 cells in which each gene is detected in scRNA-seq data. The dark blue dots mark the marker genes for each cell type (found by MAST, details in Methods). Red dots stand for the genes whose measurement scores larger than 95% percentile. Purple dots represent genes that meets both conditions. (C) Proportion of top-ranked (above the given percentile, x axis) genes that are not trivially expressed by the target cell type (defined as expression prevalence less than 20% in the target cell type) for each metric (color legend, top).

Extracellular genes and ligands are enriched by SPER detection.

If the genes spatially related to cell type composition are indeed participating in paracrine interactions, then we reasoned they must encode proteins capable of performing these functions in facilitating intercellular communication. We therefore tested whether the putative paracrine transcripts identified by SPER, and other methods were statistically enriched for (a) genes known to be secreted into the extracellular space and (b) genes known to act as paracrine ligands (**Fig. 4**). As before, we set thresholds on the metrics, requiring them to have a SPER or other metric score larger than the 95th percentile of all genes. As expected, transcripts identified by SPER were significantly enriched (p<0.001) for extracellular genes and ligands for all ten cell types. The other two metrics were significantly enriched (p<0.05) only in three and four out of ten cell types for Pearson and Spearman correlation, respectively for extracellular genes (**Fig. 4A**), and in four and three respectively for known paracrine ligand genes (**Fig. 4B**). With few exceptions (L6b neurons), SPER identified more of both types of transcripts compared to Pearson and Spearman correlations (**Fig. 4**).

Together, these results suggest that candidate paracrine genes identified by SPER have three important characteristics: (1) many are expressed by a neighboring cell type, not the target cell, but they are nonetheless spatially related to the target cell's composition, and they are also statistically enriched for (2) genes known to be secreted into the extracellular space, and (3) genes known to participate as ligands in paracrine interactions.

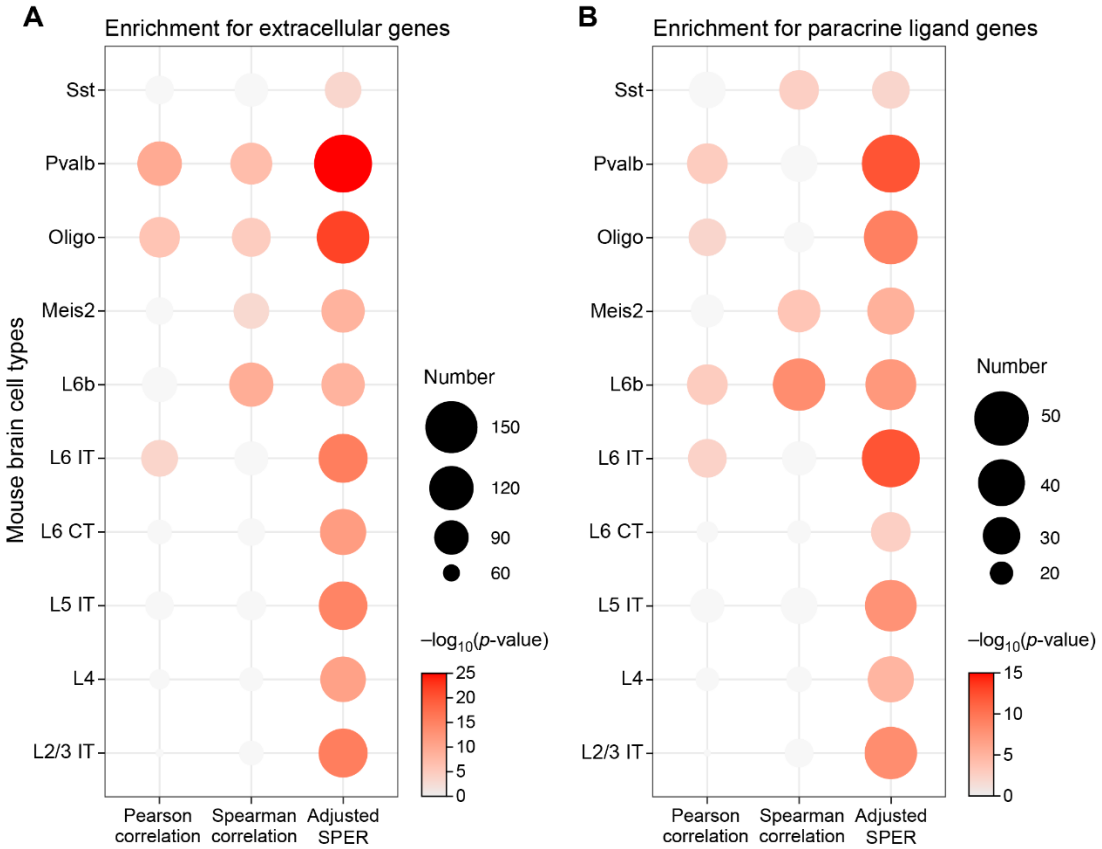


Figure 4. SPER's top-ranked genes are enriched for both extracellular genes and known paracrine ligands. Dot plots visualize hypergeometric enrichment analysis of extracellular (A) and known paracrine ligand signals (B) within candidate gene sets. Extracellular and ligand genes annotations were collected from COMPARTMENTS [25], FANTOM5 [26] and CellPhoneDB [27] databases respectively (see **Methods**). The candidate sets include genes whose scores for each spatial dependency metric (x-axis) are above than the 95% percentile of all genes. The size of dots represents the number of extracellular genes or ligand genes in the given gene set. The color represents the significance the enrichment in the set (hypergeometric test), non-significant (p>0.05) overlaps are shown as light grey.

SPER detects single- and multi-source putative paracrine signals in mouse brain.

Using SPER, we have detected a set of putative ligand genes that have high spatial association with the composition of a putative target cell type. Across all cell types there are 5569 pairs of cell types and putative paracrine regulators above the 95th percentile of SPER scores. Of these, 250 also showed substantial expression of the cognate receptor in the target cell types: a total of 506 ligand-receptor pairs, each potentially active in regulating the composition of the brain tissue in the sample. For example, *Cbln1* (precerebellin) and Meis2^{+} neurons (a cell type defined by expression of the homeobox transcription factor *Meis2* [23, 28]) showed a strong spatial association, but also non-overlapping spatial distribution, indicating possible paracrine compositional regulation (**Fig. 5B**). Notably, it was not identified by Pearson correlation, ranking 349th (R=0.32). On the other hand, using SPER we found that the *Cbln1*-Meis2^{+} neurons pair had the 4th highest score among all Meis2^{+} neuron-related signals where the appropriate cognate receptor was also expressed. Consistently, the scRNA-seq data showed that *Cbln1* is only lowly expressed in Meis2^{+} neurons, but is indeed highly expressed by the neighboring L6 IT neurons (**Fig. 5A**). In this case, reference scRNA-seq data showed that the paracrine signal is expressed only by this single source cell type. We examined *Nrxn2*, a receptor for *Cbln1* (**Fig. 5C**), and found as expected that it was expressed on Meis2^{+}

neurons, supporting the possibility of Cbln1 acting on them. In the cerebellum, *Cbln1* is known to play an important role in the Purkinje cell synapse formation [29] as well as the growth and guidance of axon in multiple neural regions within mouse embryos [30], but to our knowledge, its role as a paracrine signal in the motor cortex has not yet been described.

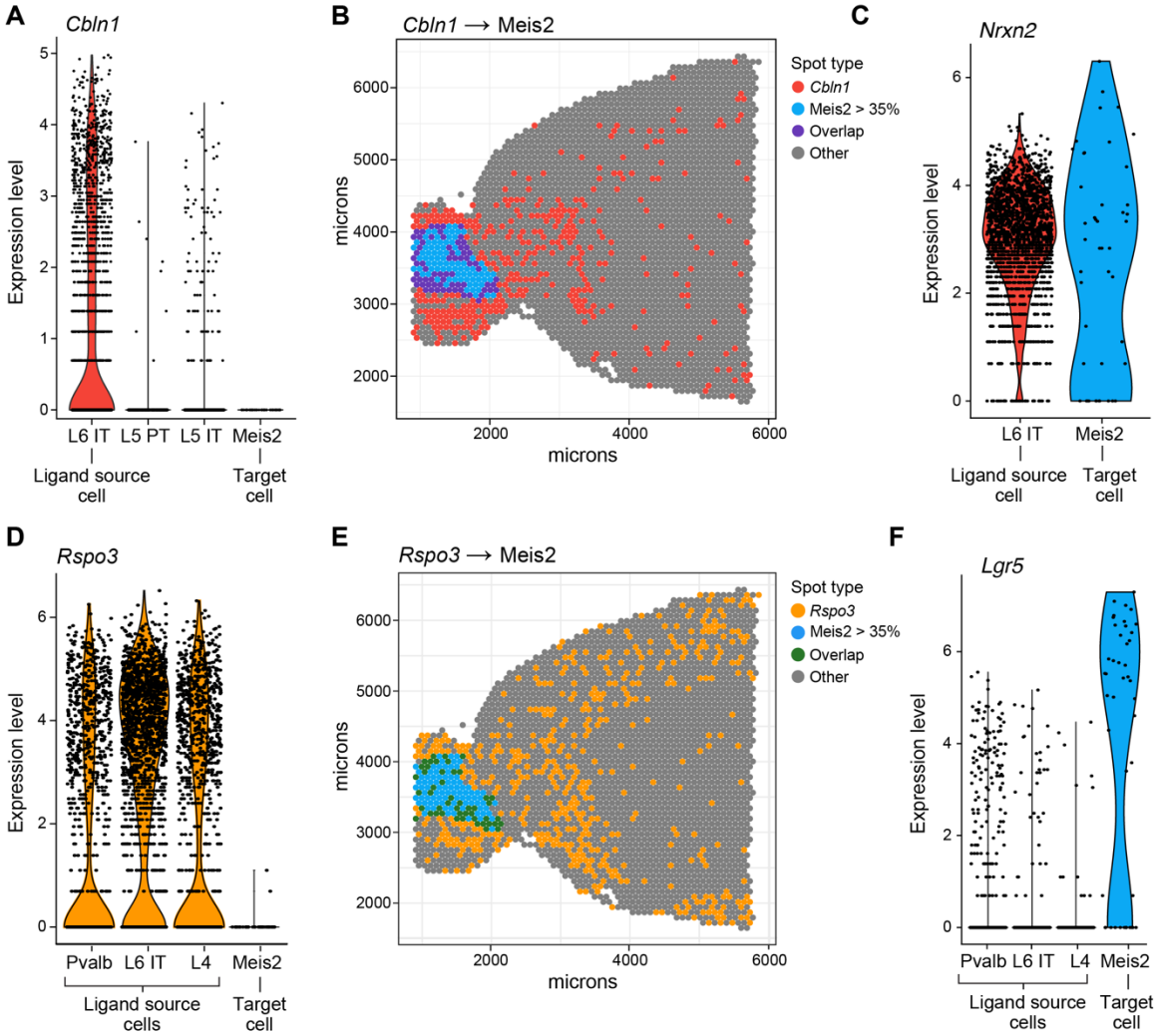


Figure 5. SPER detects single-source (*Cbln1*) and multi-source (*Rspo3*) paracrine signals in mouse brain. (**A, D**) Violin plots show expression levels (y-axis) of *Cbln1* and *Rspo3* in relevant mouse brain cell types (x-axis) from reference scRNA-seq data. L6 IT, L5 PT, L5 IT, L4, and Pvalb neurons are the neighboring cell types of Meis2$^+$ neurons. (**B, E**) ST plots visualize the spatial dependence between transcript-target cell type pairs *Cbln1*/Meis2$^+$ neurons (B) and *Rspo3*/Meis2$^+$ neurons (E). The color of spots (color legend) shows whether *Cbln1* (B) or *Rspo3* (E) is detected, a proportion of Meis2 cell larger than 35%, both (overlap), or otherwise. (**C, F**) The expression levels of the relevant cognate receptor *Nrxn2* (C) and *Lgr5* (F) in the reference scRNA-seq data.

Paracrine signals can of course also be derived from multiple cell types. For example, SPER identified *Rspo3,* which encodes the Wnt ligand R-Spondin 3. *Rspo3* is produced by several source cells: Pvalb$^+$, L6 IT, and L4 neurons (**Fig. 5D**), and shows a non-overlapping spatial association with the identified target cells, again Meis2$^+$ neurons (**Fig. 5E**). While *Rspo3* is identified as a significant putative signal by SPER, it cannot be found as a paracrine signal using any correlation metrics (**Supplementary Table 1**). We tested for expression of the cognate receptor for *Rspo3* in Meis2 cells and found that *Lgr5* (Leucine-rich repeat-

containing G-protein coupled receptor 5) was strongly and selectively expressed in the target cell type, consistent with an active paracrine interaction (**Fig. 5F**). This suggests that *Rspo3*, released by Pvalb, L6 IT and L4 cells, may serve as a paracrine regulator to Meis2 cells, activating the *Wnt* signaling pathways, which could potentially shape the spatial compositions of Meis2^{+} neurons. While this interaction is well-studied in the intestine [31], to our knowledge its role in the mouse motor cortex is not yet known. The link between *Rspo3* and Meis2 cells, therefore, provides a representative example of multi-source paracrine compositional regulators that can be discovered using SPER.

DISCUSSION:

Spatial transcriptomics is undergoing rapid development in sequencing technologies as well as computational methods. Despite the importance of compositional changes, which are frequently the first analytical question asked in ST and single-cell studies, and the potential of ST data to map them, to our knowledge there has been no method to detect putative compositional regulators; SPER provides the first such method. While ST does not have single-cell resolution, using SPER requires estimating cell type compositions from ST data using deconvolution, but the choice of decomposition methods is not a critical issue: SPER models the compositionality of data using co-kriging, reducing the impact of noise and outliers in the background while preserving core spatial patterns. But still, a method with higher decomposition accuracy will increase the performance of SPER to capture biologically meaningful signals. Hence, we expect SPER to be more powerful in the future as improved ST technologies and superior deconvolution algorithms emerge.

Though we evaluated SPER using 10X Genomics Visium data, there is nothing specific about this modality and SPER naturally generalizes to other spot-level technologies like Slide-seq [9] and XYZeq [32]. In both situations, the workflow of SPER remains the same, requiring the ST expression data which is a genes-by-location matrix; the discretized distance matrix between locations; and estimated cell-type spatial compositional data.

Algorithms for paracrine signal inference such as SPER face two related challenges given the complexity of receptor ligand interactions. First, if an interaction is previously described, and SPER identifies it as highly significant in the data, since it is based on transcriptional measurements, this is only an inference, rather than definitive proof that the interaction is physiologically relevant. In this context, SPER operates as a discovery tool, generating hypotheses regarding the paracrine interactions that could be active in the tissue, and follow-up experiments are then necessary to translate these potential interactions into definitive new biological mechanisms. As we've shown, even a cursory inspection of the results of analysis by SPER in a well-known tissue (mouse brain), identifies both known and novel potential interactions. We thus expect that SPER can provide the first step in an analytical pipeline to discover active intercellular circuits that underly tissue structure, and if applied to perturbations or disease, that underly tissue transformation. A second concern is that not all paracrine interactions are currently annotated as such – for this reason SPER does not ignore transcripts unannotated as ligands or receptors but rather enables a user to examine all highly ranked transcript-cell type pairs. In this way SPER can potentially uncover previously unknown signals.

METHODS:

SPER: Spatial Paired Expression Ratio. In general, SPER algorithm requires the normalized spatial transcriptomics data $E_{n\times m}$ (the expression level of genes divided by their global average across all spots), the spots distance matrix $M_{n\times n}$, and the cell-type compositional data $C_{k\times n}$ as the input, where n is the number of spots, m as the number of genes and k as the number of cell types. As the first step, the distance matrix of spots M is separated into l adjacency matrices A_i, each recording whether two cells lie within certain distance range $D_{i-1} \sim D_i$ for $i \in [0, l] \cap N^+$ where $D_0 = 0$. The default bin of the distance discretization is the distance between two spots, and all distances which are larger than the given largest distance limits are trimmed to that value. These adjacency matrices $A_i \in \text{Boolean}^{n\times n}$ are normalized by multiplying a diagonal matrix $F_{i,n\times n}$, where each element is the reciprocal of each row's sum of A_i. Then, they are multiplied with the spatial transcriptomics data $E_{n\times m}$ and the cell-type compositional data $C_{k\times n}$ as:

$$M \approx \sum_{i=1}^{l} D_i A_i$$

$$F_i = \text{diag}(\{\frac{1}{\sum_{k=1}^{n} A_{i,j,k}}\}_{j=1}^{n})$$

$$R_i = C \times A_i \times F_i \times E$$

where R_i represents the $k \times m$ paired expression ratio matrix at the distance range $D_{i-1} \sim D_i$. Finally, a weight function $\mathbf{\Phi} = (\varphi_1, \dots, \varphi_l)$ is applied on each gene/cell type pair across $R_1, \dots, R_l$. This weight function can follow either a step function or a Poisson distribution $\text{Pois}(\bar{\varphi})$, both of which has an extra penalty on overlapping and distal regions. Eventually, we can get the SPER score matrix S as:

$$S = \sum_{i=1}^{l} \varphi_i R_i = \sum_{i=1}^{l} \varphi_i C A_i F_i E$$

Spatial ST data simulation. The simulations are created by the following steps: 1) A 80x80 grids is created and its distance matrix is calculated; 2) Cell type labels are assigned to each tile of the grid, which follows different mechanisms in three scenarios; 3) For each cell type, we create the corresponding marker genes and simulated their expression levels across all cells. Corresponding cell types will have higher counts of their markers with their counts following Poisson distributions $\text{Pois}(\lambda)$ where $\lambda \sim \text{Unif}(15, 30)$. For other genes, including the non-marker and the control genes, their counts follow Poisson distributions $\text{Pois}(\lambda')$ where $\lambda' \sim \text{Unif}(0, 3)$. 4) The compositional data for each cell type is computed from marker genes' expression level. The count of each cell type within a grid is sampled from another Poisson distribution. The final cell-type proportions are the fractions of each cell types within the tiles. Mathematically it can be shown as:

$$\tilde{y}_{g,k} \sim \text{Pois}(\lambda_{g,k})$$

$$c_{i,k} \sim \text{Pois}(\sum_{g \in \text{marker of i}} \alpha \lambda_{g,k} \tilde{y}_{g,k})$$

$$\tilde{c}_{i,k} = \frac{c_{i,k}}{\sum_i c_{i,k}}$$

$$\lambda_{g,k} \sim \begin{cases} \text{Unif}(15, 30), & \text{if } g \text{ is a marker} \\ \text{Unif}(0, 3), & \text{otherwise} \end{cases}$$

where $\tilde{y}_{g,k}$ is the expression level of gene g at tile k, $\tilde{c}_{i,k}$ is the cell-type proportion of cell type i at tile k, and α is the scale factor.

Simplified single-source scenario simulates the source and target cells in a spatial relationship as ring (**Fig. 2**). Within the 80x80 grid, five cell types, two pairs of source and target cells and a control, are simulated. For each simulated cell type, we created 20 marker genes from two-dimensional normal distributions and calculated the cell-type spatial composition as described above. While the target cells have their marker genes simulated with given means and variances, the markers for source genes have their maximal values 10-unit away from the centers of target cell types. This spatial dependence characterizes the paracrine patterns between the source genes and target cells.

Realistic scenarios, both single-source and multi-source, we use Markov random field algorithm from the R package 'mrf2d' to simulate the spatial relationships between cells. Neighbors with L1 distance less than 3 are considered when calculating the 2D Markov random field. Source and target cells are set to spatially dependent on each other, while the target cells have a greater spatial autocorrelation. The single-source scenario simulated one pair of source and target cells and two control cell types, while the multi-source having two sources, one target, and two control cell types. The Pearson and Spearman correlation, and SPER scores are respectively calculated on three simulations. The resulting matrices are used to predict the spatial dependence between genes and cell types, particularly the target cells and source genes. For each type, the ROC curve is plotted in distinguishing its markers and paracrine signals from the other genes.

Collection and pre-processing of scRNA-seq reference data. The scRNA-seq data were collected and pre-processed by Tasic, B., et al. in the Allen Institute for Brain Science, generated by the SMART-seq2 protocol [23]. The dataset contains 14249 adult mouse brain cortical cells and 34617 genes. The raw data was normalized by Seurat using SCTransform function [33], where 3000 subsampling cells were used to build the NB regression model and other parameters set as the default. The label 'subclass' was used as the annotation of cell type, which included 23 different subclass of mouse brain cells. Cajal–Retzius (CR) cell type, the count of which was only 7 in the scRNA-seq data, was removed from the data as the low cell counts weakened its statistical power. After this removal, 14242 cells with 22 distinct cell types were kept as the reference in the downstream analysis.

Identifying cell-type-specific marker genes in scRNA-seq data. The cell-type-specific marker genes were found in the normalized scRNA-seq data by Seurat. For each cell type, their marker genes were then identified by Seurat using MAST one-versus-all test [24]. In these tests, each gene was tested whether it was differentially expressed in each cell type compared to rest of cells. Genes are selected as marker genes by the following criteria: fold change > 2, adjusted p-value < 10^{-30}, fraction of expressing cells > 70%. All the other parameters were set as the default in the 'FindMarkers' function in the Seurat package [33].

Collection and pre-processing of Visium spatial transcriptomics data in mouse brain. The mouse brain spatial transcriptomics data is collected from the 10X Genomics Visium public database [8]. It contains 2696 spots with 124367 mean reads per spot. The raw count matrix is normalized by Seurat using SCTransform function with default settings. The physical distances of cells are calculated from the provided spatial coordinates and the scale factor in the metadata. In the downstream analysis where ST and scRNA-

seq data are needed, we only kept the genes detected in both datasets and having the total UMI counts greater than 40 in the normalized ST data.

Cell-type deconvolution and co-kriging modeling in Visium mouse brain data. In the Visium mouse brain example, we used the normalized ST data to perform the cell-type deconvolution. The cell-type spatial compositional data C is estimated using RCTD full mode [10] with a reference using the scRNA-seq reference data. We select top 10 cell types which have the highest prevalence among all spots in the compositional data generated by RCTD. They are L2/3 IT, L5 IT, Meis2, L4, L6 IT, Oligo, Pvalb, L6 CT, Sst, and L6b. The other cell types are summed together as 'Other' aside from those 10 types. The cell-type spatial compositional matrix is further denoised using the co-kriging algorithms. 15 distance bins, 120 μm each, are used to compute the log-ratio variograms. To model the variograms between cell types, we use a compositional linear model of spatial coregionalization that includes a nugget effect term and 3 Gaussian variogram terms whose respective effective as 600, 1300, and 3000 micrometers. These 3 terms respectively represent the short-range, middle-range, and long-range interactions between cell types.

SPER calculation. In the Visium mouse brain example, we used the normalized ST data and cell-type spatial compositional data as the SPER inputs. Calculation was performed using SPER R package. The weight function is derived from a Poisson distribution $(\mathrm{Pois}(2.5) - 0.05) \times 5$ to give penalty on overlapping and distal regions while keeping its sum as 1. The adjusted SPER score was calculated as:

$$\tilde{s}_{i,j} = \frac{s_{i,j}}{\exp\,(p_{i,j} - 1)}$$

Where $s_{i,j}$ stands for the SPER score and $p_{i,j}$ for the expression prevalence of gene i and cell type j.

Collection of extracellular gene set and ligand-receptor pair data. To identify the subcellular locations, we used the all channels integrated version of mouse extracellular gene set that includes knowledge, experiments, text mining, and predictions channel from the COMPARTMENTS database [25]. To filter out the intracellular genes, we select the genes whose locations are in the extracellular region (including extracellular space, matrix, exosome, and vesicle). Also, we only kept the genes whose confident level greater than 3. As a result, 4797 piece of genes' information and 2622 different genes are left as the set of extracellular genes. After doing the intersection with the ST data, 1784 genes are included in the final extracellular gene set. The ligand-receptor interaction dataset is collected from FANTOM5 ligand-receptor datasets [26] and CellPhoneDB database [27]. Since all these databases only have the human genome data, we transformed the gene ID into the homologous gene in mouse genome to compare with the ST data we have got. There are totally 1888 ligand-receptor pairs that have the homologous genes in mouse, where 1636 of them are from FANTOM5 and 252 of them from CellPhoneDB.

Calculating correlations and EMD of gene-cell type pairs. The Pearson and Spearman correlations are calculated between the spatial compositional data after co-kriging modeled and the spatial transcriptomics data. The same matrices, but size reduced, are used to calculate the EMD in comparing the computing complexity. We use the physical distance between spots as the ground distances, and the proportions of each cell-type composition are inputted as the weights when comparing to a gene's spatial distribution. R package 'emdist' is used to calculate EMD in the simulations.

Spatial-dependent signals detection. The SPER score matrix, Pearson correlation matrix and Spearman correlation matrix are used to filter the candidate genes and spatial-dependent signals with the help of scRNA-seq data as the reference of expression levels. When studying the enrichment of extracellular and ligand genes, we used the hypergeometric test to see if they are enriched within the gene set given thresholds. In both cases, the population size is 13436; the number of successes in the population is 1784 for the extracellular and 1888 for ligands. The sample size and the number of successes in the sample are dependent on the expression-level and measurement-level thresholds. If the p-values calculated from the tests are less than 0.05, we will take it as significantly enriched.

Acknowledgements

The authors thank Leslie Gaffney for preparing and generation of figures for this article, Evan Lemire for programming assistance, and Yu Sun for helping improve the data visualization.

Author contributions

T.Z. designed and implemented the SPER algorithm, performed data analysis and visualization and wrote the manuscript. A.H. conceived the study and wrote the manuscript.

Declaration of interests

The authors declare no competing interests.

Data and Code availability

An R package to implement the SPER algorithm is available for download at github.com/TianxiaoNYU/SPER. Extracellular gene set in mouse is available from the COMPARTMENTS database: https://compartments.jensenlab.org/Downloads; scRNA-seq reference data is obtained from Allen Institute for Brain Science: https://portal.brain-map.org/atlases-and-data/rnaseq; Visium spatial transcriptomics data from mouse brain is available from 10X Genomics dataset: https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-0-0. An R markdown tutorial to help generate the main results of this article using SPER package, as well as all other necessary data files in this study can be found at: https://www.dropbox.com/scl/fo/pep07jjid71rr72voeeik/h?rlkey=2agyojo1ugf5l61bqtal6hjy0&dl=0.

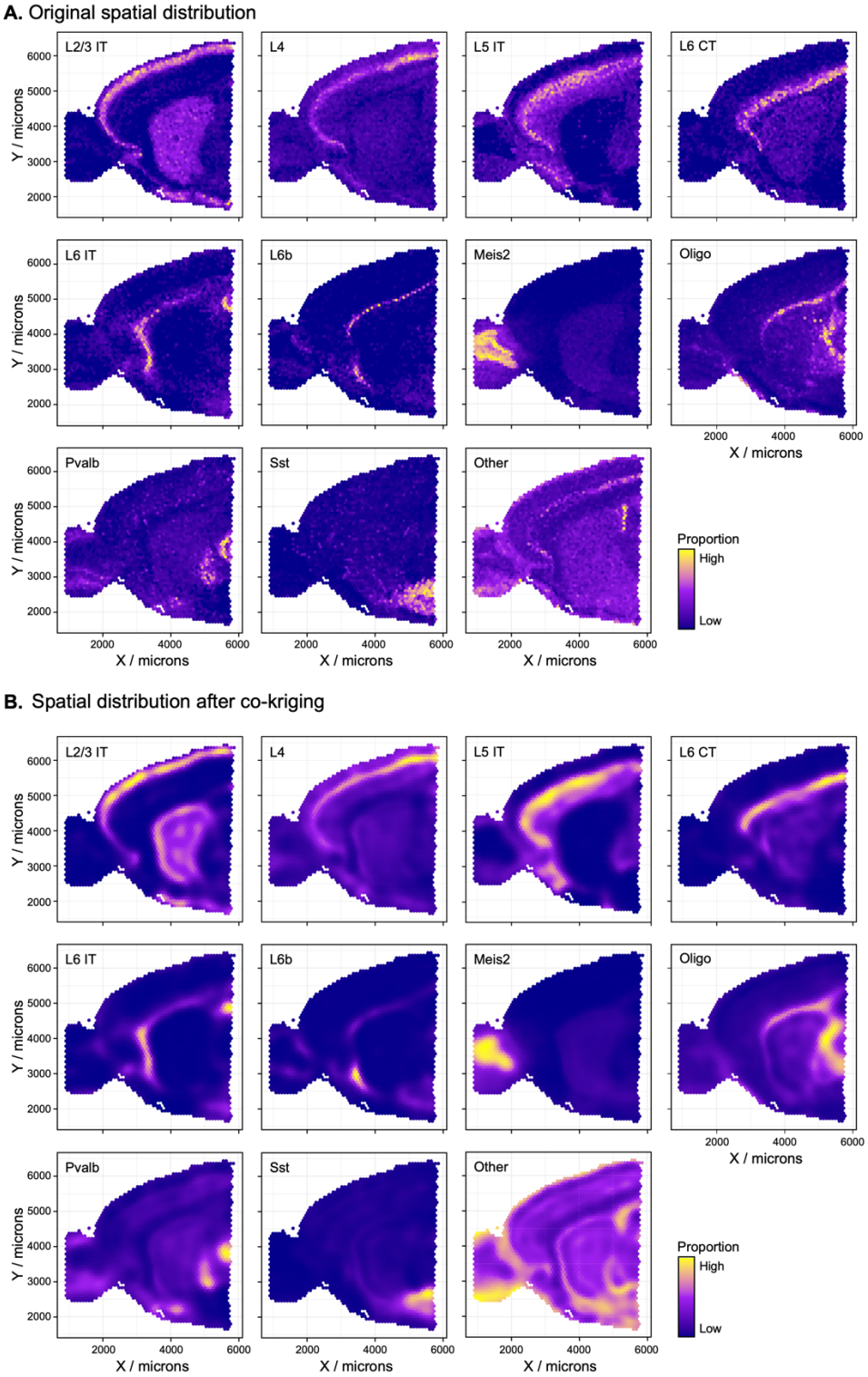
REFERENCES:

1. Anthony, R.M., et al., *Protective immune mechanisms in helminth infection.* Nature Reviews Immunology, 2007. **7**(12): p. 975-987.
2. Johnson, T.S., et al., *Combinatorial analyses reveal cellular composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer's Disease.* Sci Rep, 2021. **11**(1): p. 353.
3. Pawelec, P., et al., *The Impact of the CX3CL1/CX3CR1 Axis in Neurological Disorders.* Cells, 2020. **9**(10).
4. Knoop, K.A. and R.D. Newberry, *Goblet cells: multifaceted players in immunity at mucosal surfaces.* Mucosal Immunol, 2018. **11**(6): p. 1551-1557.
5. Carr, E.J., et al., *The cellular composition of the human immune system is shaped by age and cohabitation.* Nat Immunol, 2016. **17**(4): p. 461-468.
6. Ansell, S.M. and R.H. Vonderheide, *Cellular composition of the tumor microenvironment.* Am Soc Clin Oncol Educ Book, 2013.

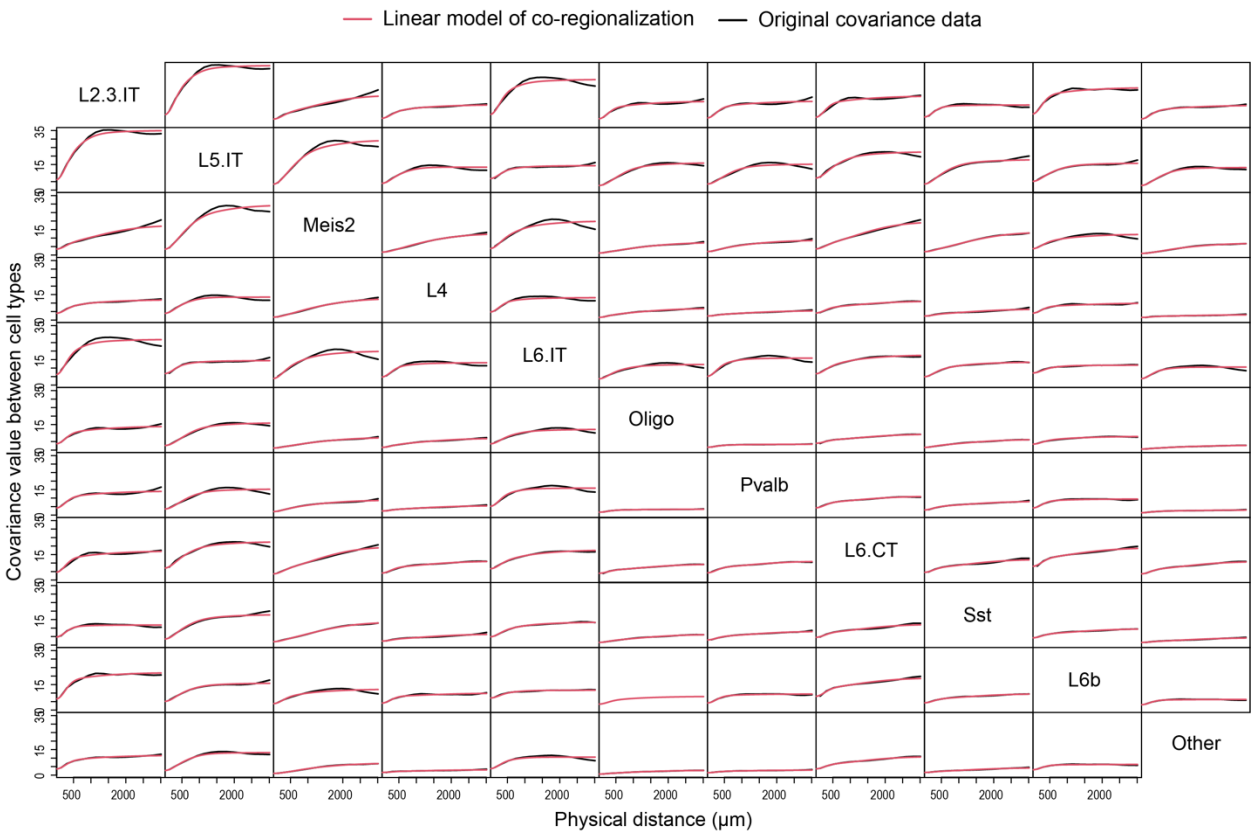
7. Palla, G., et al., *Spatial components of molecular tissue biology.* Nat Biotechnol, 2022. **40**(3): p. 308-318.
8. Genomics, X. *Mouse Brain Serial Section 1 (Sagittal-Anterior)*. 2019 [cited 2021 10/9]; Available from: https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-0-0.
9. Rodriques, S.G., et al., *Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution.* Science, 2019. **363**(6434): p. 1463-1467.
10. Cable, D.M., et al., *Robust decomposition of cell type mixtures in spatial transcriptomics.* Nat Biotechnol, 2022. **40**(4): p. 517-526.
11. Kleshchevnikov, V., et al., *Cell2location maps fine-grained cell types in spatial transcriptomics.* Nat Biotechnol, 2022. **40**(5): p. 661-671.
12. Liu, Z., et al., *SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics.* Nat Commun, 2023. **14**(1): p. 4727.
13. Tuong, Z.K., et al., *Resolving the immune landscape of human prostate at a single-cell level in health and cancer.* Cell Rep, 2021. **37**(12): p. 110132.
14. Park, J., et al., *Cell segmentation-free inference of cell types from in situ transcriptomics data.* Nat Commun, 2021. **12**(1): p. 3545.
15. Grisanti Canozo, F.J., et al., *Cell-type modeling in spatial transcriptomics data elucidates spatially variable colocalization and communication between cell-types in mouse brain.* Cell Syst, 2022. **13**(1): p. 58-70 e5.
16. Baker, E.A.G., et al., *In silico tissue generation and power analysis for spatial omics.* Nat Methods, 2023. **20**(3): p. 424-431.
17. Tolosana-Delgado, R., U. Mueller, and K.G. van den Boogaart, *Geostatistics for Compositional Data: An Overview.* Mathematical Geosciences, 2019. **51**(4): p. 485-526.
18. Satoh, A., *Chapter 10 - Typical Properties of Colloidal Dispersions Calculable by Molecular-Microsimulations*, in *Studies in Interface Science*, A. Satoh, Editor. 2003, Elsevier. p. 153-159.
19. Myers, D.E., *Matrix Formulation of Co-Kriging.* Journal of the International Association for Mathematical Geology, 1982. **14**(3): p. 249-257.
20. Krige, D.G., *A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand.* Journal of the South African Institute of Mining and Metallurgy, 1994. **94**(3): p. 95-111.
21. Rubner, Y., C. Tomasi, and L.J. Guibas, *A metric for distributions with applications to image databases.* Sixth International Conference on Computer Vision, 1998: p. 59-66.
22. Freguglia, V. and N.L. Garcia, *Inference Tools for Markov Random Fields on Lattices: The R Package mrf2d.* Journal of Statistical Software, 2022. **101**(8): p. 1 - 36.
23. Tasic, B., et al., *Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.* Nature Neuroscience, 2016. **19**(2): p. 335-+.
24. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biology, 2015. **16**.
25. Binder, J.X., et al., *COMPARTMENTS: unification and visualization of protein subcellular localization evidence.* Database-the Journal of Biological Databases and Curation, 2014.
26. Ramilowski, J.A., et al., *A draft network of ligand-receptor-mediated multicellular signalling in human.* Nature Communications, 2015. **6**.
27. Efremova, M., et al., *CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.* Nature Protocols, 2020. **15**(4): p. 1484-1506.
28. Frazer, S., et al., *Transcriptomic and anatomic parcellation of 5-HT3AR expressing cortical interneuron subtypes revealed by single-cell RNA sequencing.* Nature Communications, 2017. **8**(1): p. 14219.
29. Ito-Ishida, A., S. Okabe, and M. Yuzaki, *The role of Cbln1 on Purkinje cell synapse formation.* Neurosci Res, 2014. **83**: p. 64-8.
30. Han, P., et al., *Cbln1 regulates axon growth and guidance in multiple neural regions.* PLoS Biol, 2022. **20**(11): p. e3001853.

547 31. Barker, N., et al., *Identification of stem cells in small intestine and colon by marker gene Lgr5.*
548 Nature, 2007. **449**(7165): p. 1003-7.
549 32. Lee, Y., et al., *XYZeq: Spatially resolved single-cell RNA sequencing reveals expression*
550 *heterogeneity in the tumor microenvironment.* Sci Adv, 2021. **7**(17).
551 33. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902
552 e21.
553
554
555
556

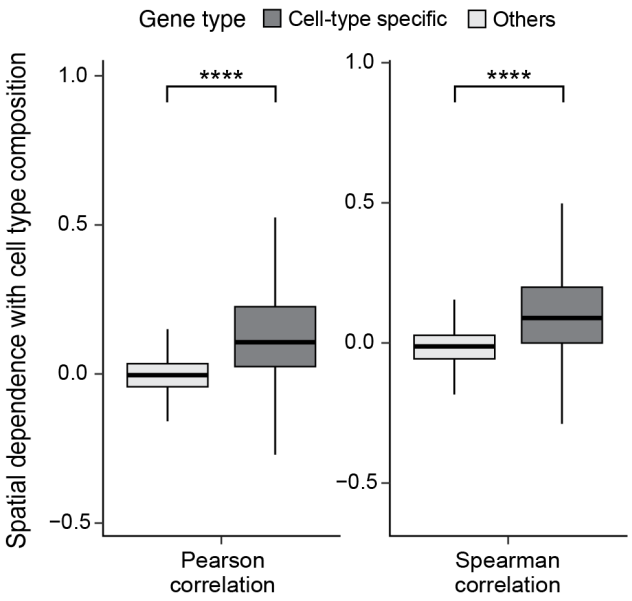
557 **SUPPLEMENTARY FIGURES:**



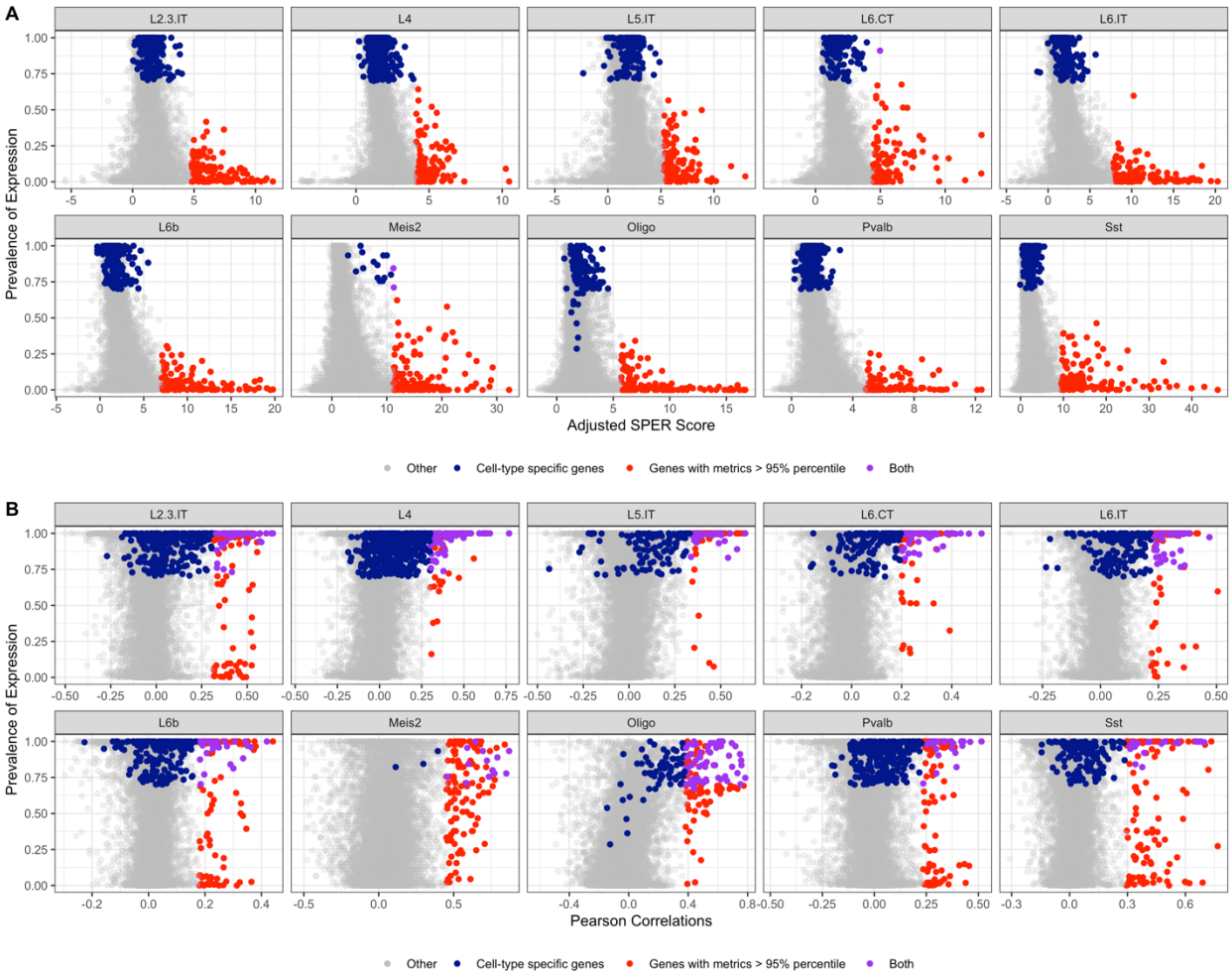
558
559 **Supplementary Figure 1. Co-kriging enables SPER to effectively denoise spatial compositional data**. Spatial distribution of
560 composition of primary cell types decomposed by RCTD in full mode before (A) and after co-kriging modeling (B). The top ten
561 most abundant cell types are modeled individually, and the other 11 cell types are summed as the 'Other' category. In each
562 subplot, X and Y-axis represent the physical coordinates of spots, and the color represents the normalized proportion of the
563 relevant cell type.



Supplementary Figure 2. Co-kriging fits visualized as variograms. Co-kriging method fits on the ten neuronal cell types from an example dataset from mouse brain. A compositional linear model of co-regionalization (red line) with a 'nugget' effect term for short scale randomness and three Gaussian variogram terms to model effects at different ranges (600, 1300, and 3000 µm) showed excellent fit to the original covariance data (black line) collected and calculated at 15 uniformly distributed lags. The labels at the diagonal identify the cell type for each row and column, while 'Other' represents the sum of all remaining cell types. The x-axis in each subpanel represents the physical distance in µm, and y-axis represents the relative value of the covariance between each pair of cell types.



575
576 **Supplementary Figure 3. Pearson and Spearman correlations detect marker genes with overlapping spatial dependency.**
577 Boxplots showing Pearson and Spearman correlations between cell type spatial compositions and gene's spatial distribution.
578 Marker genes for each cell types are selected from the scRNA-seq reference data by MAST. The four-star labels represent the
579 difference (adjusted p-value < 10^{-4}) in the measurement scores between the marker and non-marker groups.



Supplementary Figure 4. Comparison of trivial spatial associations and genuine putative regulators between SPER and Pearson correlation. Scatter plots show the distribution of SPER (A) and Pearson correlation (B) scores (x axis) with the proportion of cells of each type (plots) in which each gene (points) is detected (y axis) in scRNA-seq data. The dark blue dots mark the marker genes for each cell type (found by MAST, details in Methods). Red dots stand for the genes whose measurement scores larger than 95% percentile. Purple dots represent genes that meets both conditions.