Discovering paracrine regulators of cellular composition from spatial transcriptomics data using SPER

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

[1] Institute of Systems Genetics, NYU Grossman School of Medicine, New York, NY

[2] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA.

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

ABSTRACT:

A wide range of pathologies and chronic diseases are characterized by transformations in cell type composition within a tissue. Prominent examples are remodeling of barrier epithelia in response to external threats, immune cell infiltration in chronic inflammatory conditions, or proliferation of fibroblasts and stromal cells causing fibrosis. An extensive literature implicates paracrine signals secreted by the local tissue niche in regulating its cell type composition, and therefore dysregulation of these factors is frequently capable of inducing pathological changes. Emerging multi-omics methods, particularly spatial transcriptomic profiling, provide the ability to simultaneously assay cell type composition and gene expression, with the potential to discover the transcripts most strongly associated with compositional shifts. However, no method yet exists to detect such signals. Here, we develop Spatial Paired Expression Ratio (SPER), a computational approach to evaluate the spatial dependence between transcripts and cell types 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 the adult mouse motor cortex, we show that genes identified by SPER are strongly statistically enriched for both transcripts known to be extracellularly secreted, and those known to participate in paracrine receptor-ligand pairs, providing evidence of the ability to detect compositional regulatory signals. SPER recovered known and novel interactions, such as a surprising spatial association between the Wnt ligand *Rpso3* and its cognate receptor *Lgr5*, expressed in a specific neural subset, and 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; Cellular Composition; Spatial Compositional Data Analysis; Cell-cell communication; Systems Biology

38 **INTRODUCTION:**

39 To gain mechanistic into the pathophysiology of tissue transformations associated with disease, it is
40 essential to understand what signals affect the compositions of cell types in tissue. Many diseases involve
41 this type of change in corresponding tissues: in Alzheimer's disease, research has found that the cellular
42 composition changes are associated with the transcriptomic changes of cell-type-specific genes including
43 many key genes and pathways involved [1]; as another example, in asthma, airway remodeling happens in
44 certain areas where it is narrowed by tightened muscles and increased numbers of mucus-producing
45 epithelial cells (goblet cells) accompanied by the abnormal emerging of immune cells [2]. Triggering
46 factors, like environmental exposures [3] and microenvironment signals [4], can induce cellular
47 composition changes. A complete description of the biological mechanisms driving these changes remains
48 elusive, but essential to answering this question is to find potential biological signals that initiate this
49 process.

50

51 Recent studies applying single-cell genomics have shown that even rare types of cells can play important
52 roles in triggering compositional change by producing the paracrine signals to recruit inflammatory cells
53 [5]. A famous example of such signals in the asthmatic airway is IL-13 (Interleukin-13) in type 2 immune
54 response where it is secreted by the Th2 cell is the key regulator that induces the replacement of epithelial
55 cells with goblet cells [6]. This well-studied signal is in fact the central mediator of allergic asthma that
56 connects the cellular signal pathway, cellular composition changes, and allergic response together.

57

58 Spatial transcriptomics (ST) is an overarching term for a range of methods designed to capture
59 transcriptome-wide gene expression while retaining information about the spatial tissue context [7].
60 Therefore, unlike transcriptome sequencing methods like bulk or single-cell RNA sequencing (scRNA-seq),
61 spatial transcriptomics provides us access to analyze cells' physical locations, offering the potential to map
62 out cell-to-cell spatial interaction and communication [8]. The resolution of different technologies for
63 spatial transcriptomics varies from multi-cell level to subcellular level. Particularly, the most widely-used

and relatively affordable technologies, including ST Visium [9] and Slide-seq [10], capture the spatial transcriptomics information with a spot-based protocol typically of diameter 55 microns, thus losing the capability to recover single-cell level expression patterns.

Given the recent emergence of this data modality, there is a relative paucity of computational tools for analyzing spatial transcriptomic data, and in particular, to our knowledge there are no methods designed for the analysis of cell-type compositional data, which has a unique set of characteristics requiring specialized analytical methods [11]. Currently, the method to address spatial compositional analysis uses enrichment scores as a metric for the abundance of each cell type across the space, which does not model the compositional nature of the relative abundance between cell types [12]. As statisticians going back to Pearson have pointed out, this can lead to the problem of *spurious correlations* [13], which may translate in this instance to spurious transcript-cell type abundance interactions.

The major analytical field to derive inferential tools for spatial compositional data (CoDa) has been geostatistics, a branch of statistics for modeling data that show codependency related to their spatial proximity, originally developed in predicting the distribution of metal ores for mining [14]. Developed in the last century, principles and theories have been developed in this field. Co-kriging, as a powerful tool in analyzing the spatial CoDa, can model the compositional relevance as well as the spatial correlations between components. It uses the correlation function which describes the pattern of the covariance by a pair of components with geographically a lag distance apart [13]. This method enables us to analyze the cell-type spatial CoDa with proper principles and avoid the spurious correlation brought by the constant sum of components if we just treat it as other ordinary spatial data.

The first step towards spatial compositional analysis from spatial transcriptomic data to is to *deconvolve* the aggregate gene expression signal to estimate the relative proportions of component cell types, a problem currently well addressed by several published approaches. Seurat [15] applies an anchor-transferring

algorithm to transfer the cell type labels from a reference scRNA-seq dataset to the spatial transcriptomics data. Two other prominent methods decompose the spot-level expression data into the cell type abundances using each cell type's distinctive expression pattern. Spatial Dampened Weighted Least Squares (SpatialDWLS) [16], part of the Giotto toolkit [12], uses weighted least square deconvolution to estimate cell type proportions based on the expression level of previously defined cell type marker genes. Robust Decomposition of Cell Type (RCTD) [17], uses an analogous approach, this time based on maximum likelihood estimation. All these methods produce data which has the constraint that in each spot all proportions sum to 1, and thus their output is *compositional*. A general problem when analyzing this type of regulator signal, however, is to determine whether the signals and the receptor cell types are sufficiently spatially close to each other that the signaling process is possible. Even though there could be an overlap between the spatial distribution between signals and receptors cells, the diffusion of signal molecules must also be considered in the signal transduction process. For example, one of the TGF-β paracrine signals, Nodal acts over distances from 40 to 200 μm corresponding to approximately 15 cell diameters [18]. As a result, such spatial overlapping between signal expressions and receptor cells might disappear in the spatial transcriptomics data and cell-type compositional data.

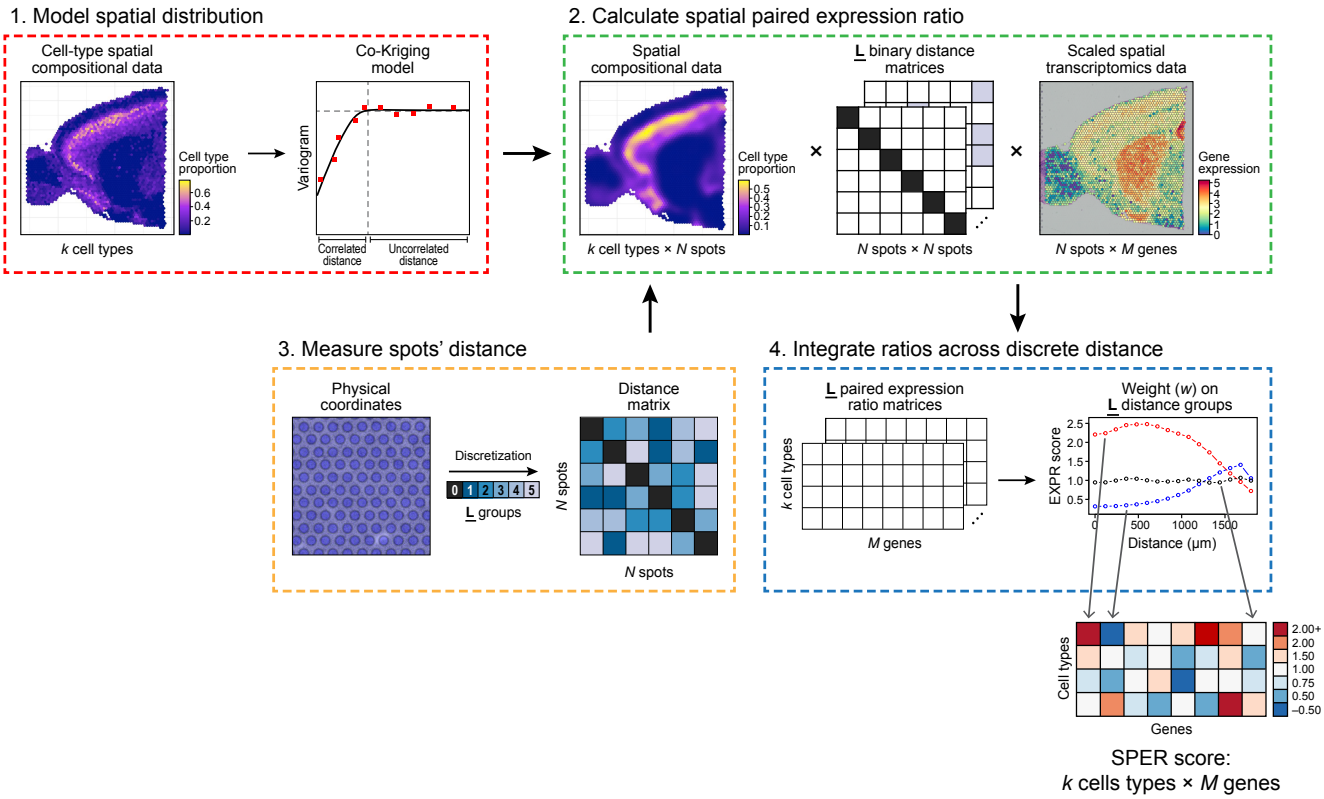


Figure 1. Schematic of spatial paired expression ratio (SPER) algorithm. The workflow of SPER: (1) cell-type spatial compositional (ST) data algorithms is denoised using co-Kriging; (2) physical coordinates, included in spatial transcriptomics data, are discretized into the distance matrix, which is then converted to several binary distance matrices; (3) combining the above outcomes and ST expression data, we evaluate the spatial association between genes and cell types, and calculate the SPER matrix; (4) finally, we apply a weight at the proposed distance groups to get the SPER score for every pair of gene and cell type.

To overcome this problem, we introduce Spatial Paired Expression Ratio (SPER), a spatial-transcriptomics-based approach to detect the paracrine signals whose target cells and expression patterns have non-overlapping spatial dependence (**Fig. 1**). The idea of SPER is inspired by the pair correlation function (radial distribution) in physics, which is to evaluate 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 [19]. To model the compositionality of our spatial data we apply co-kriging from the geostatistics field [20], 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 outperforms other applicable metrics - Pearson correlation, Spearman correlation, and Earth Mover's Distance (EMD) - on simulated datasets. We next apply SPER to ST data of adult mouse brain and show that SPER is able to more accurately identify a set of candidate paracrine regulators compared to the other metrics – this is assessed by showing candidate genes are more significantly enriched for both extracellular genes and known paracrine ligands. Finally, we show that SPER can identify novel, biologically significant interactions, such as the precerebellin (*Cbln1*) associated with Meis2 neurons (which express its receptor *Nrxn2*), and the Wnt ligand R-spondin 3 (*Rspo3*) associated with the same Meis2 neuron subset (which express its receptor *Lgr5*). Both putative paracrine signal links are identified by SPER but cannot be distinguished by the correlation scores. 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.

133 **RESULTS:**

134 **RCTD shows high accuracy in estimating cell type proportions from ST data.**

135 To make compositional inferences from ST data, our first step is to estimate cell-type proportions by
136 deconvolving spot-level spatial transcriptomics data. We use 10X Visium ST data from mouse brain
137 anterior [9] as our testing dataset. To compare the performance of different cell type
138 decomposition/deconvolution methods, we trained a random forest classifier as a gold standard in
139 predicting the largest cell-type component of each spot in the ST data, which showed high accuracy (OOB
140 estimate of error rate as 0.88%). We then applied and compared all previously published spatial cell type
141 decomposition methods, 'anchor'-based integration in Seurat v3 [15], RCTD [17], and SpatialDWLS [16],
142 to get the cell type spatial compositional data from spatial transcriptomics data. All these methods require
143 normalized scRNA-seq data as a reference to learn and then predict the cell type compositions in each ST
144 spot. Two different modes of Seurat anchor transferring, PCA-based (principal component analysis) and
145 CCA-based (canonical correlation analysis), are separately tested; three different modes of RCTD, doublet
146 'mode' which assumes the maximum number of cell types in a spot is 2, 'full' mode which assumes no
147 limitation on the cell type number, and 'multi' mode which uses a greedy algorithm to determine the number
148 of cell types. We reasoned that the most accurate spatial deconvolution algorithms should show the highest
149 agreement with the most common cell type estimated by the random forest classifier, which we report as
150 accuracies (**Fig. S1**).

151 We first tested the running time of each method to estimate cell type proportions on our test dataset, which
152 contained 2695 spots and 1118 genes (S1 Fig). We chose the overall accuracy and multi-class AUROC
153 (Area Under ROC) as the metrics, which modified the binary AUC concept to fit the multi-class problem
154 and usually had a higher value than the accuracy since it generally started from 0.5 to 1. This result is
155 consistent with the published articles that RCTD has a better accuracy than spatialDWLS and Seurat while
156 having a slower computing speed [16]. The overall accuracies and the multi-class AUROCs (area under the
157 receiver operating characteristic) of these predictions show that RCTD in full mode has the best
158 performance with the highest accuracy and AUROC both across all spots and across a set of 'purer' spots,

159 the top 10% of spots whose primary components have the largest compositions (**Fig. S1**). As a negative
160 control, we randomly shuffled the predicted labels for all spots and compared them with their original one,
161 which, as expected showed a very low accuracy and a multi-class AUROC of approximately 50%.
162 Therefore, after considering both aspects, we selected RCTD in full mode as the cell-type compositional
163 method for the following analyses.
164

165 **SPER captures simulated spatial relationships between paracrine signals and composition of target**
166 **cell type.**

167 The SPER algorithm is described in detail in the methods. Briefly, it first applies co-kriging algorithms to
168 model the cell-type spatial compositional data; then, using the distance information obtained from the
169 physical coordinates and the scaled spatial transcriptomics, it calculates the paired expression ratio as a
170 series of matrices, which includes all pairs of cell types and genes; finally, depending on the interaction
171 distance people are interested in, SPER adds a customized weight on ratios to eventually obtain one score
172 representing the spatial relationship between the given cell type and the gene. To evaluate its performance,
173 we first applied SPER to a simulated dataset (see Methods) and compared it with traditional methods to
174 detect co-variation, specifically: Pearson correlation, Spearman correlation, and Earth Mover's Distance
175 (EMD) [21]. We simulated two classes of signaling genes: 'overlap' signals, whose spatial distributions
176 overlap with their corresponding cell types; and 'paracrine' signals, whose spatial distributions are
177 dependent on the cell types' but not necessarily overlapped (**Fig. S3**). The 'overlap' class of genes
178 corresponds to the trivial case where a transcript's spatial abundance is associated with the composition of
179 a cell type because *that cell type is the exclusive source of that transcript*. The second 'paracrine' class of
180 genes, represent putative paracrine interactions, where the *transcript is secreted by a neighboring cell type*
181 and then potentially acts on the target cell type located nearby. We simulated paracrine gene expression by
182 a source cell type at a fixed distance R from the target cell type (S3 Fig, Methods). Clearly, accurate
183 detection of genes in the 'paracrine' class is the goal of SPER and our study.
184

185 We found that unlike classical correlation methods, SPER accurately detected signals in both scenarios (**Fig.**
186 **2A**). Its average AUROCs in both two cases are near 1, meaning it can accurately identify the relationship
187 between both types of signals and their matched cell types. While SPER performance was slightly lower
188 than correlations when detecting the overlapping signals, its AUROC was 0.981, still indicating excellent
189 performance. Traditional correlations (Pearson and Spearman) can identify the overlapping signals but were
190 substantially worse (both AUROC around 0.75) in detecting the paracrine ones. This result can be explained
191 by the nature of correlation functions: they ignore the spatial information and only measure the similarity
192 of corresponding values. Therefore, since the paracrine signals have a *spatially similar but not directly*
193 *corresponding distribution* to their target cell types, they cannot be captured by these traditional methods.
194 We have also considered the Earth Mover's Distance (EMD), which is a measurement of spatial distance
195 between two distributions over a region [21]. Unlike the correlations, EMD accurately captures both types
196 of signals in the simulation data as effectively as SPER. However, the computing speed of EMD is much
197 slower than SPER: even in the simulation data where the number of spots is small (500 spots), EMD is
198 approximately 100 times slower than SPER when evaluating 10 pairs of genes and cell types (**Fig. 2B**).
199 This difference becomes even larger when the grids become larger (**Fig. 2C**), which makes it impractical
200 to apply EMD on the real ST data (usually greater than 2000 spots). As the last part of our simulation, we
201 want to test if SPER still outperforms other measurements in a more complicated situation, where the
202 distance between paracrine signals and target cell types is allowed to vary. Here, we adjusted the distance
203 parameter R from 1 to 10 units and then evaluated the performance of SPER, Pearson or Spearman
204 correlations, as EMD was impractical to run (**Fig. 2D**). As expected, SPER can detect the signals well even
205 when they are more away from the targeting cell types compared to Pearson or Spearman correlations,
206 whose AUROC values decline significantly with the increase of distance (i.e, decreasing overlap). Taken
207 together, our simulation studies demonstrate the ability of SPER to detect paracrine signals that might be
208 missed by other traditional metrics.
209

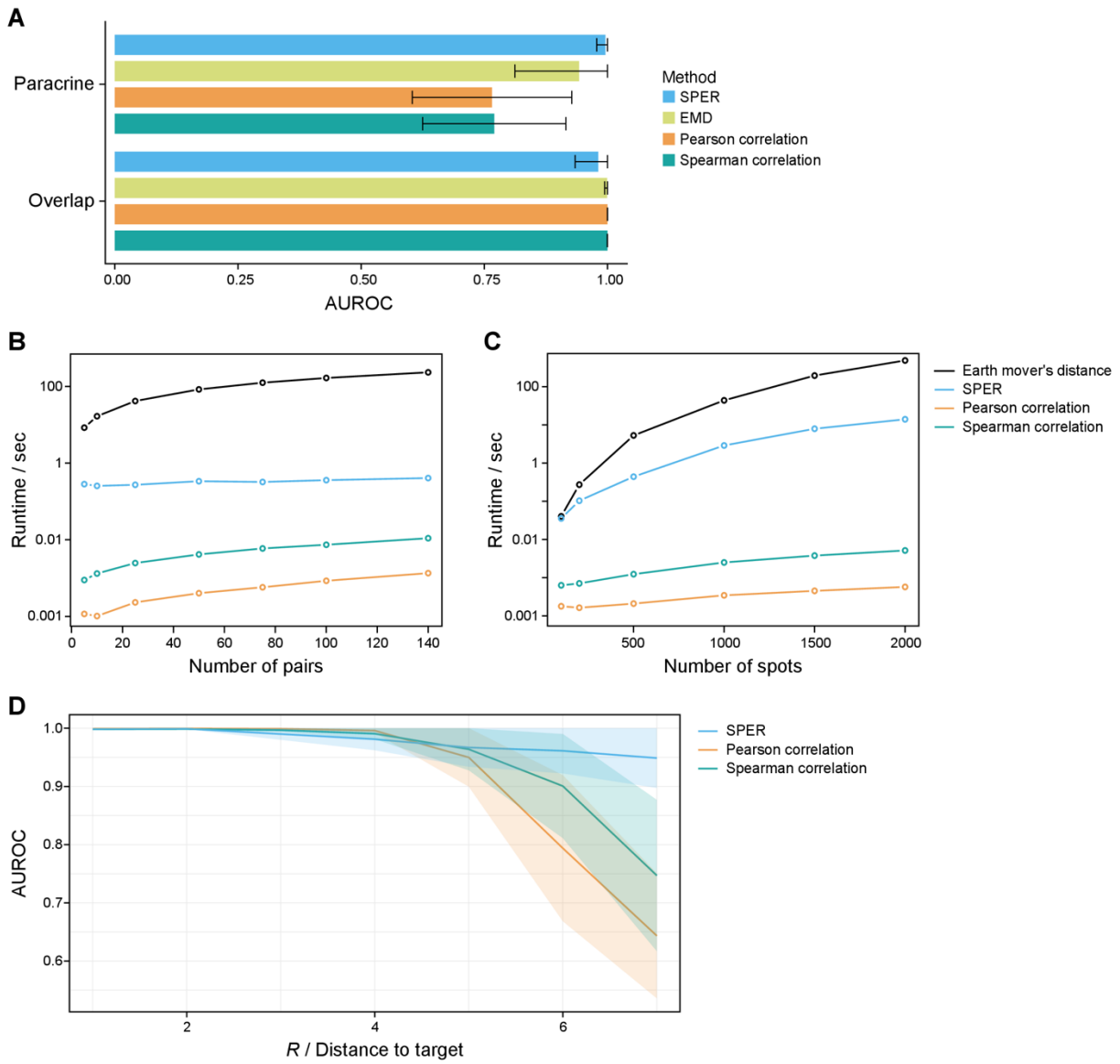


Figure 2. Evaluation of SPER on simulated spatial transcriptomics data. (A). The bar plot showing the AUROC of SPER, Earth mover's distance, Pearson, and Spearman correlations at detecting the spatial signals in the simulations. Paracrine signals are set to be 7.5 units away from the cell type centers. The bar represents the average AUROC of a method in detecting a type of signal. The lengths of error bars stand for the standard error of the AUROC under 50 simulations, and those values that exceed 1 are trimmed to 1. (B, C) The runtime in seconds comparison between methods in the simulations. Each method is tested within a 500-spot grid with a different number of randomly generated gene/cell-type pairs in B, and within the different shapes of grids with 10 gene/cell-type pairs in C. (D) The average AUROC at

218 detecting the paracrine signals based on SPER, Pearson or Spearman correlations. Different distances to
219 the target cell type centers, from 1 to 8 units, were applied to the signals in the simulations. The lines in
220 different colors represents the average AUROCs under given signal distance evaluated from different
221 measurements. Shallow areas represent 95% confidence intervals.
222
223 **Co-kriging models the compositionality of estimated spatial cell-type proportions.**
224 Co-kriging, a variant of kriging methods which models the regionalized data to interpolate the spatial
225 distribution, is widely applied in the analysis of spatial compositional data as it can effectively models the
226 covariance between the multiple regionalized variables that are spatially related to each other [13]. The first
227 step in the SPER algorithm (**Fig. 1**) is therefore an application of co-kriging to properly model the spatial
228 relationship between each cell type. The model (described in Methods) accurately estimates the covariance
229 between the proportions of *n* cell types as well as the residual variance (**Fig. S2**). This result suggests the
230 method can not only model the desired inter-cell-type spatial relationship, but also remove background
231 noise (**Fig. 3** and **S4**). Except where otherwise noted, all downstream analyses use this co-kriging-modeled
232 estimates of cell type proportions.

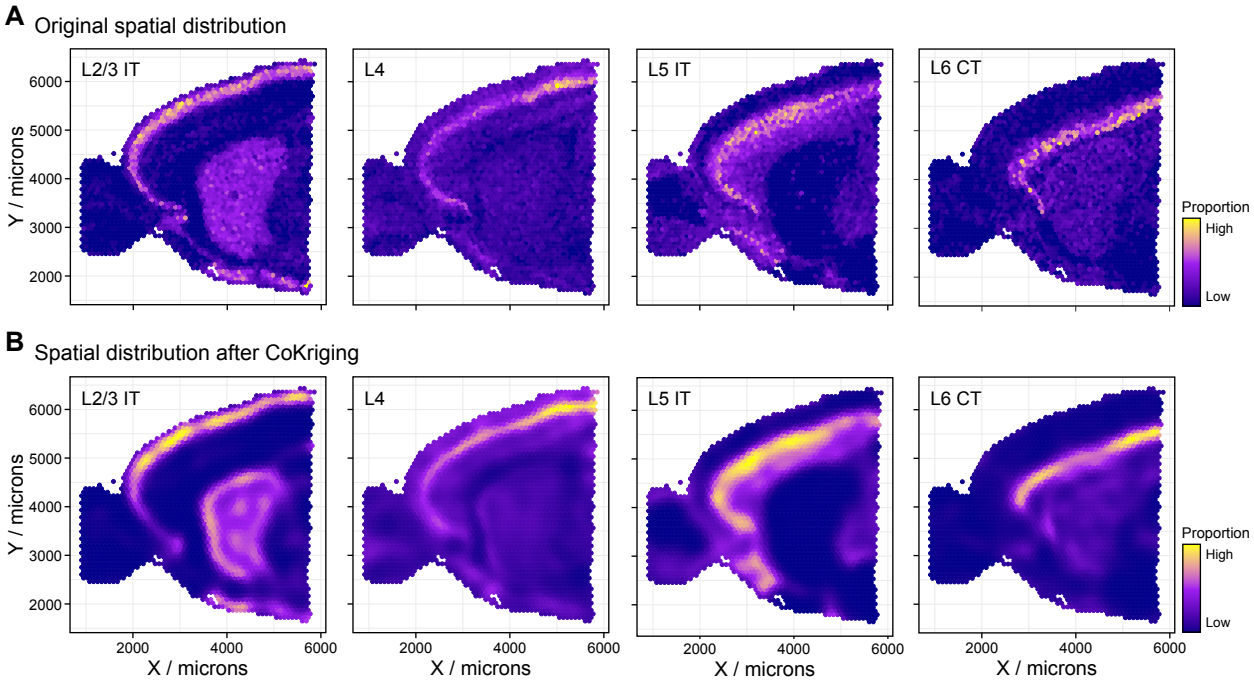
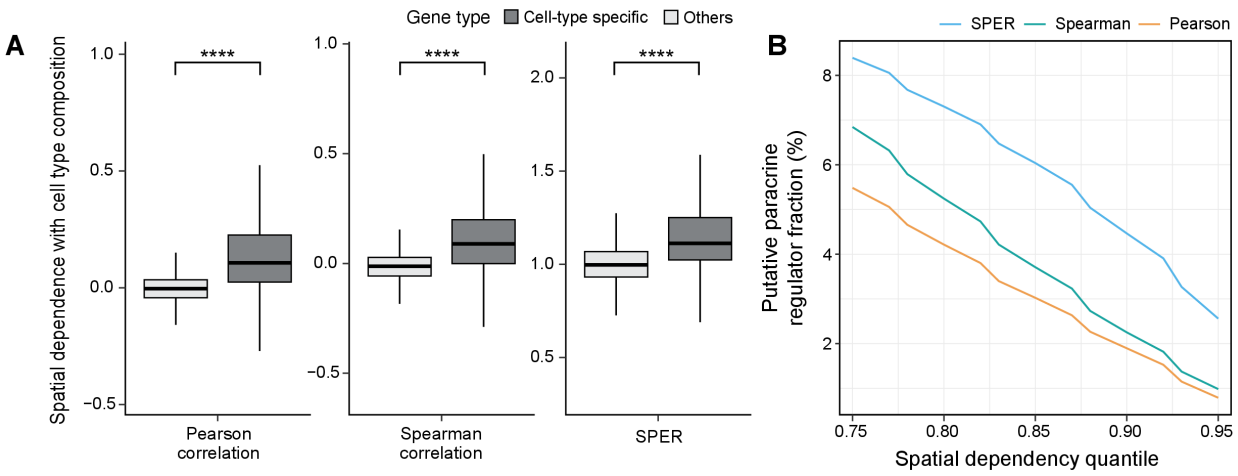


Figure 3. Co-kriging models effectively denoise spatial compositional data. We visualized the spatial compositional data of the 4 primary cell types (L2/3 IT, L4, L5 IT, and L6 CT) decomposed by RCTD full mode before (A) and after co-kriging (B). In each subplot, X and Y-axis represent the physical coordinates of spots in μm, and the color represents the compositional proportion from low to high.

SPER enables detecting marker genes and candidate genes.

To evaluate its performance on real data we next applied SPER, along with the other metrics, to Visium ST data from mouse brain [9], first estimating cell type proportions in each spot using RCTD and then using co-kriging to simultaneously model compositionality and smooth out noise in the corresponding spatial compositional estimates data (**Fig. 3**, **Methods**). For a given cell type *A*, 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 sole source of these transcripts; this is analogous to the 'overlap' genes described above. Far more interesting, from our point of view, are the candidate 'paracrine' genes, which show a spatial relationship with the composition of *A but are not expressed* by *A* cells.

248 For each of the ten neuronal cell types annotated in our test dataset, we first confirmed that cell-
249 type-specific genes, defined as genes having a significant differential expression in one particular cell type
250 (adjusted p-value < 1e-30, $|\log_2(\text{Fold-change})| > 1$, calculated by MAST [22]), were – trivially, as described
251 in the preceding paragraph – spatially related to the composition of that cell type, as expected (**Fig. 4A**).
252 More importantly, since our goal is to identify paracrine signals, we expect SPER to detect those genes that
253 are not expressed, or only lowly expressed, in the a given cell type but do have expression in the other cell
254 types that are spatially nearby, which we name 'putative paracrine signals'. For each cell type-gene pair we
255 compared the proportion of cells of that type expressing that gene from matched scRNA-seq data [23] with
256 each metric for spatial relationship detection (Pearson, Spearman, and SPER, **Fig. S5**). We defined
257 candidate paracrine genes as those whose expression prevalence was less than 20% in the given cell type
258 and varied the different percentile threshold for the spatial relationship detection metric (**Fig. 4B**). Under
259 this criterion, SPER identified substantially more genes than either Pearson or Spearman correlation
260 (approximately twice at the 90% percentile threshold), confirming its capability to detect putative paracrine
261 regulators of cell type composition.
262



263

264 **Figure 4. SPER detects both putative paracrine signals and cell-type specific marker genes.** (A)
265 Boxplots showing Pearson correlation, Spearman correlation, and SPER scores distribution of cell type-
266 specific 'marker' and other genes. Marker genes for each cell types are selected from the scRNA-seq

267 reference data by MAST (details in Methods). The four-star labels represent the significant difference
268 (adjusted p-value < 10^{-4}) in the measurement scores between the marker and non-marker groups. (B) The
269 fraction of candidate genes whose expression prevalence is less than 20% in the paired cell type and the
270 measurement scores are larger than the given percentile of all genes' scores. The X-axis stands for the given
271 percentile and Y-axis for the fraction in percentage.
272
273 **Extracellular genes and ligands are enriched by SPER detection.**
274 To further validate our hypothesis, we tested whether these 'candidate paracrine genes' identified by SPER
275 are statistically enriched for (a) genes known to be secreted into the extracellular space and (b) genes known
276 to act as paracrine ligands. We also set thresholds on the metrics levels for those candidates, requiring them
277 to have a SPER or correlation score larger than the 95% percentile of all genes, to study the enrichment
278 situations in each primary cell type. The result suggests that the extracellular genes and ligands are more
279 enriched in certain cell types like Oligo, Meis2, and L6 IT (**Fig. 5**). We have also examined the results in
280 Pearson and Spearman correlation data, whose enrichments are not as significant as SPER: given the 95%
281 percentile of SPER scores and 20% expression prevalence threshold, the extracellular genes are more
282 significantly enriched in most primary cell types (except L2/3 IT, L4, and Pvalb cells) compared to Pearson
283 and Spearman correlation under this threshold (**Fig. 5A**). Also, the numbers of extracellular genes filtered
284 by the SPER score in each significant cell type are larger than that by Pearson or Spearman correlation.
285
286 We observed similarly improved performance for detection of known paracrine ligand genes, where for
287 most primary cell types (except L4, L6 CT, and Pvalb cells) SPER candidate genes are more significantly
288 enriched with the ligands under the same threshold compared to Pearson and Spearman (**Fig. 5B**). Also, the
289 number of ligands captured by SPER in those significant cell types is greater than that of the correlations.
290 This result suggests that SPER is a better method than the correlations in detecting the candidate signals
291 which have spatial-dependent expression patterns but are rarely expressed in the matched cell type. Taken
292 together, these results suggest that candidate paracrine genes detected by SPER have three important

characteristics: (1) they are expressed by a neighboring cell type, not the target cell, but they are 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.

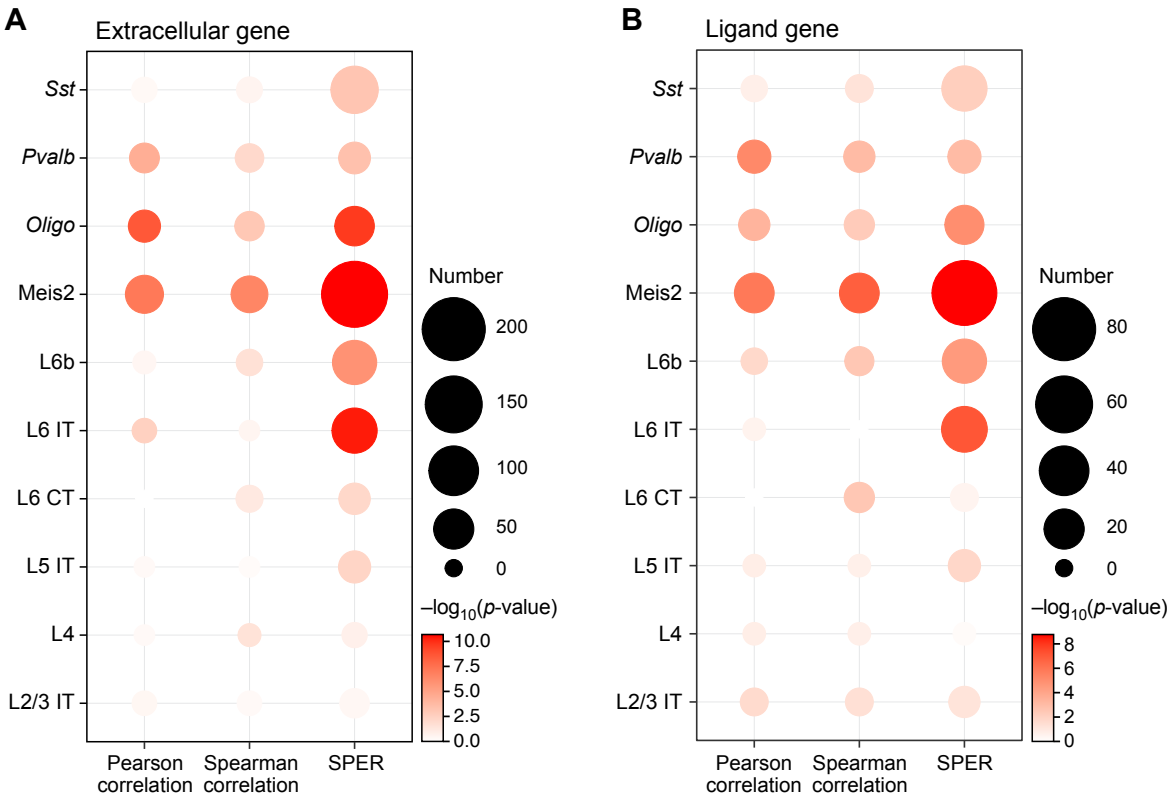


Figure 5. SPER's top-ranked genes are enriched for both extracellular genes and known paracrine ligands. The enrichment analysis of extracellular (A) and ligand signals (B) within candidate gene sets. Extracellular and ligand genes annotations were collected from COMPARTMENTS [24], FANTOM5 [25]and CellPhoneDB [26]database (see **Methods**). The candidate gene sets include genes whose expression prevalence less than 20% in their targeting cell types (y-axis), and measurement scores (x-axis, as SPER or correlations) larger than the 95% percentile of all genes'. In each subplot, the size of dots stands for the number of extracellular genes or ligand genes in the given gene set. The red color represents the significant level of the enrichment in the set.

307

308 **SPER detects single-source (*Cbln1*) and multi-source (*Rspo3*) paracrine signals in mouse brain.**

309 Using SPER, we have detected a set of putative ligand genes that have high spatial association with the
310 composition of a putative target cell type. There are 344 putative paracrine regulators (2.56%) in the 95th
311 percentile of SPER scores, of these, 96 also showed substantial expression of the cognate receptor in their
312 target cell types, a total of 137 signal/target pairs. All these 96 genes are potentially active compositional
313 regulators in the tissue. For example, *Cbln1* (precerebellin) and Meis2 neurons (a cell type defined by
314 expression of the homeobox transcription factor *Meis2* [23] [27]) showed a strong spatial association (most
315 possible signaling distance around 300 μm), but also few overlaps, indicating possible paracrine
316 compositional regulation (**Fig. 6B**). Consistently, the scRNA-seq data showed that *Cbln1* is only lowly
317 expressed in Meis2 cells but is indeed highly expressed by the neighboring cell type L6 IT neurons – a
318 potential regulator (**Fig. 6A**). We term this type of regulatory interaction 'single-sourced' since the
319 paracrine signal is expressed only by a single cell type, L6 IT neurons in this case. As a further validation,
320 we examined the expression profile of *Nrxn2*, one of *Cbln1* receptors, in our reference data (**Fig. 6C**).
321 Contrary to the ligand, Meis2 displayed relatively expression of its receptor *Nrxn2*, supporting the
322 possibility of Cbln1-acting on Meis2 neurons. In the cerebellum, *Cbln1* plays an important role in the
323 Purkinje cell synapse formation [28] as well as the growth and guidance of axon in multiple neural regions
324 within mouse embryos [29], but to our knowledge, its role as a paracrine signal in the motor cortex has not
325 yet been described.

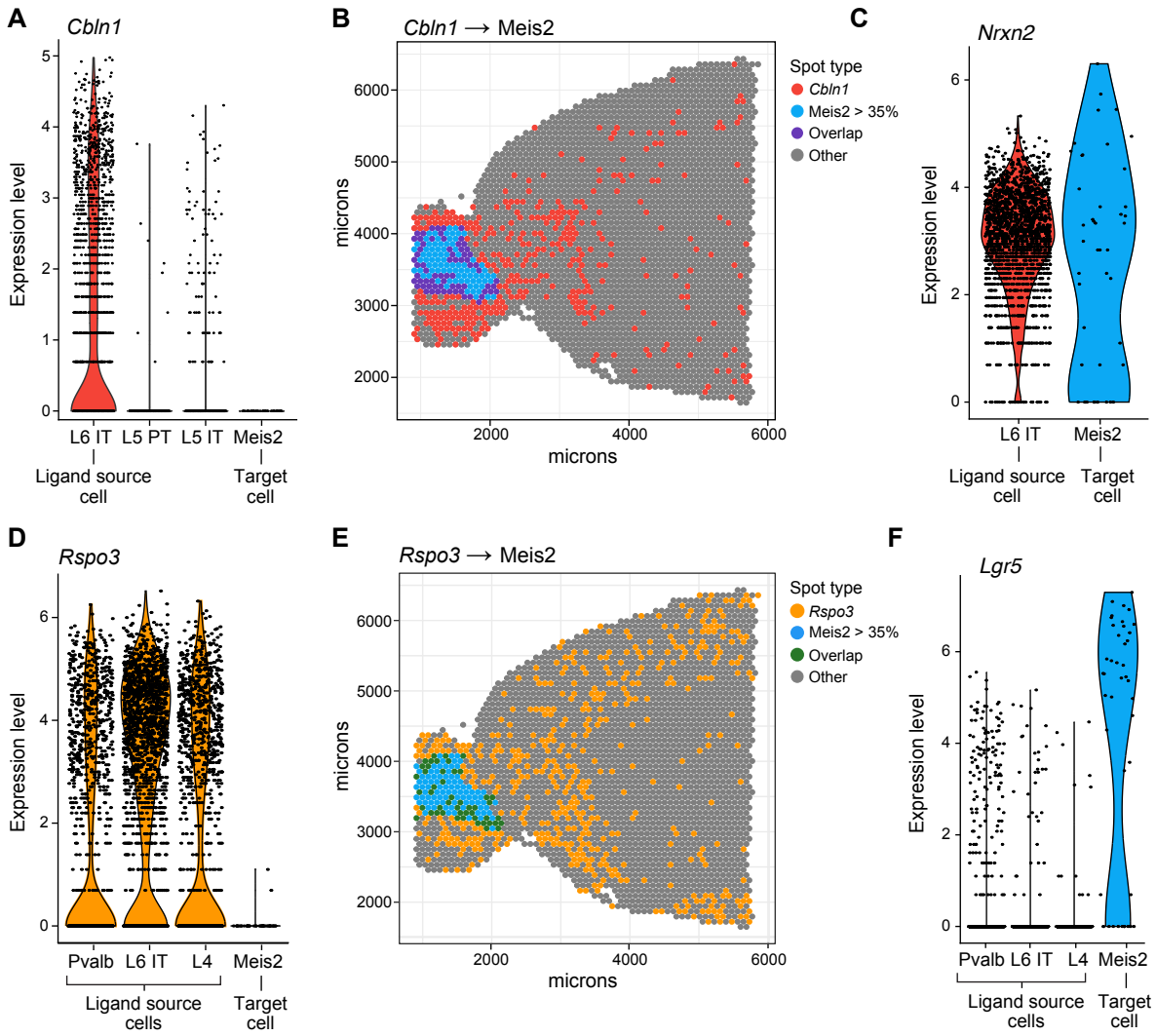


Figure 6. SPER detects single-source (*Cbln1*) and multi-source (*Rspo3*) paracrine signals in mouse brain. (A, D) The expression levels of *Cbln1* and *Grp* in the reference scRNA-seq data. The X-axis stands for different cell type annotated in the reference scRNA-seq data, and the Y-axis stands for the normalized expression levels in each single cell. L6 IT, L5 PT, L5 IT, L4, and Pvalb cells are the neighboring cell types of Meis2. Among them, L6 IT has the highest expression level of *Cbln1* and Pvalb, L6 IT and L4 have highest expression levels of *Rspo3*, making them potential targeting cell types for signaling. (B, E) Spatial plots showing paracrine pattern between *Cbln1*/Meis2 pair and *Rspo3*/Meis2 pair: the color of spots shows if a spot express *Cbln1* or *Rspo3*, with a proportion of Meis2 cell larger than 35%, having both cases (Overlap), or otherwise. (C, F) The expression levels of *Nrxn2*, one of *Cbln1*'s receptor, and *Lgr5*, one of

336 *Rspo3*'s receptor, in the reference scRNA-seq data. Presented cell types and figure legends are same as in
337 panel A and D.
338
339 Conversely, paracrine signals that target a cell type can also be derived from multiple cell types. For
340 example, *Rspo3,* which encodes the Wnt ligand R-Spondin 3, is produced by several source cells: Pvalb,
341 L6 IT, and L4 cells (**Fig. 6D**), and also shows a high non-overlapping spatial association with the target
342 cells, again Meis2 neurons (**Fig. 6E**). We tested for expression of cognate receptor for *Rspo3* in Meis2 cells,
343 and found that *Lgr5* (Leucine-rich repeat-containing G-protein coupled receptor 5) was strongly and
344 selectively expressed in the target cell type, consistent with an active paracrine interaction (**Fig. 6F**). while
345 this interaction is well-studied in the intestine [30], to our knowledge its role in the mouse motor cortex is
346 not yet known. This finding indicates a significant as well as potentially important paracrine signaling
347 relation in mouse brain, that *Rspo3*, released by Pvalb, L6 IT and L4 cells, may serve as a paracrine regulator
348 to Meis2 cells by activating the *Wnt* signaling pathways which can potentially shape the spatial
349 compositions of neurons that are Meis2^{+}. The link between *Rspo3* and Meis2 cells, therefore, provides a
350 representative example of multi-source paracrine compositional regulators that can be discovered using
351 SPER.

352 **DISCUSSION:**

353 Spatial transcriptomics is undergoing rapid development in sequencing technologies as well as analytical
354 methods. Despite the importance of compositional changes, however, and the potential of ST data to map
355 them, to our knowledge there has been no method to detect putative compositional regulators; SPER
356 provides the first such method.

357

358 Typically, we expect that the most straightforward way to obtain cell type compositional data is through
359 de-convolution, and while ST methods do not yet have single-cell resolution, this will continue to be
360 necessary. Here we evaluated several published methods and found the best performance with RCTD
361 (currently as a part of spacexr); while we were preparing the manuscript, two further methods that are often
362 used and compared have emerged which we did not evaluate: CellDART [31] and DestVI [32]. Since it is
363 difficult to obtain a true 'gold standard', which is the actual spatial distribution of cell-type proportions, in
364 the spot-level spatial transcriptomics data, it is still hard to say which cell-type decomposition method is
365 the 'best' in terms of its accuracy. One possible solution is through simulation, where people can design
366 spatial patterns and generate the simulated spatial transcriptomics data with known cell-type compositions
367 [11]. Other solutions are mostly based on other spatial transcriptomics data which have a higher resolution
368 like seqFISH [16]. For SPER, however, the choice of decomposition methods is not a critical issue: with
369 co-kriging, SPER models compositionality of data, reducing the impact of noise and outliers in the
370 background while preserving the general spatial patterns. But still, a method with higher decomposition
371 accuracy will increase the performance of SPER to capture biologically meaningful signals. Hence, we
372 expect SPER to be more powerful in the future as improved ST technologies and superior deconvolution
373 algorithms emerge.

374

375 SPER operates as a discovery tool by generating hypotheses regarding the paracrine interactions that could
376 be active in the tissue, and follow-up experiments are of course necessary to translate these potential
377 interactions into definitive new biological mechanisms. However, as we've shown, even a cursory

inspection of the results of analysis by SPER in a well-known tissue (mouse brain), generate 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. Though we evaluated SPER using 10X Genomics Visium spatial transcriptomics data, there is nothing specific about this modality and we expect it to generalize to other spot-level technologies like Slide-seq [10] and XYZeq [33] and importantly, the co-kriging aspect of SPER enables it to generalize to single-cell and subcellular technologies (such as MERFISH [34]) although this will require additional cell segmentation [35]. In both situations, the workflow of SPER will remain the same provided 3 necessary input data: the ST expression data which is a genes-by-location matrix; the discretized distance matrix between locations; the cell-type spatial compositional data reflecting the cell types' spatial distribution. As discussed previously, the weights are also changeable depending on the data type. This flexibility lets SPER be a universal computing method for detecting the spatial-dependent signals in spatial transcriptomics data.

METHODS:

SPER: Spatial Paired Expression Ratio.

In general, SPER requires the scaled spatial transcriptomics data $E_{n\times m}$ (the expression level of genes divided by their global average across all spots), the spots physical coordinates $M_{n\times n}$, 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.

First, the distance matrix of spots M is calculated and discretized into l groups with a given rule which user can modify. 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:

$$M \rightarrow M_1, \dots, M_l$$

where each matrix M_i is a $n \times n$ binary matrix indicating if a distance lies within the group i. These matrices are then right multiplied by a diagonal factor matrix $F_{i,n\times n}$ whose elements are the reciprocal of each row's sum. Then, they are right multiplied by the spatial transcriptomics data E and left multiplied by the cell-type compositional data C:

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

where R_i is the $k \times m$ paired expression ratio matrix at the distance group i.

As the final step, a weight vector $\boldsymbol{w} = (w_1, \dots, w_l)$ is applied on the ratios at the same location across $R_1, \dots, R_l$, where $w_1, \dots, w_5 = \frac{1}{5}$, $w_6, \dots, w_l = 0$ as the default setting. By this integration, we can finally get the SPER score matrix S:

$$R_1, \dots, R_l \xrightarrow{\text{Weight } \boldsymbol{w}} S_{k\times m}$$

In this study, we used the scaled ST data as E, its corresponding distance matrix as M, co-kriging-modeled cell-type spatial CoDa as C, and the default weight $\boldsymbol{w}$ to calculate the SPER scores. The distance bin is 120

423 μm, which is approximately the average distance between two spots in 10X Genomics Visium sequencing
424 data.
425
426 **Collection and pre-processing of scRNA-seq reference data.**
427 The scRNA-seq data were collected and pre-processed by Tasic, B., et al. in the Allen Institute for Brain
428 Science, generated by the SMART-seq2 protocol [23]. The dataset contains 14249 adult mouse brain
429 cortical cells and 34617 genes. The raw data was normalized by Seurat using SCTransform function [15],
430 where 3000 subsampling cells were used to build the NB regression model and other parameters set as the
431 default. The label 'subclass' was used as the annotation of cell type, which included 23 different subclass
432 of mouse brain cells. Cajal–Retzius (CR) cell type, the count of which was only 7 in the scRNA-seq data,
433 was removed from the data as the low cell counts weakened its statistical power. After this removal, 14242
434 cells with 22 distinct cell types were kept as the reference in the downstream analysis.
435
436 **Identifying cell-type-specific marker genes.**
437 The cell-type-specific marker genes were found in the normalized scRNA-seq data by Seurat. For each cell
438 type, their marker genes were then identified by Seurat using MAST one-versus-all test [22]. In these tests,
439 each gene was tested whether it was differentially expressed in a given cell type compared to the rest cells.
440 A threshold on the expression level was given, that gene should be expressed in at least 70% of cells in the
441 corresponding cell type. Candidate genes whose fold change were greater than 2 and their adjusted p-value
442 less than 10^{-30} were kept as marker genes. All the other parameters were set as the default in the
443 'FindMarkers' function in the Seurat package [15].
444
445 **Collection of extracellular gene set and ligand-receptor pair data**.
446 To identify the subcellular locations, we used the all channels integrated version of mouse extracellular
447 gene set that includes knowledge, experiments, text mining, and predictions channel from the
448 COMPARTMENTS database [24]. The data includes 3046243 piece of genes' location information with

their confidence level from 1 to 5. 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 [25]and CellPhoneDB database [26]. Since all these databases only have the human genome data, we transformed the gene ID into the homologous gene in mouse genome in order 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.

Decomposition of cell type compositions.

In Seurat [15], the transfer anchors are first calculated by integrating the ST data and scRNA-seq using either PCA or canonical correlation analysis (CCA) reduction. The first 30 components are kept in both methods. The compositional data is then predicted by cell type label transferring with the help of the anchors found. In RCTD [17], we take the normalized counts in scRNA-seq data as the input and generate the reference object where the specific features for each cell types are automatically calculated by RCTD. The spatial data is also normalized by their total UMI counts in each spot before the prediction. Three different computing modes, 'full', 'doublet', and 'multi', have all been applied and evaluated. In SpatialDWLS [16], a Giotto object is first created with the spatial transcriptomics data. As the reference, we calculate the average expression level of the top 200 marker genes of each cell type based on their adjusted p-value in the scRNA-seq data. This mean-expression matrix serves then as the input to run SpatialDWLS where the number of cells per spot set as 20. All algorithms were run on the Harvard Cannon cluster using 6 cores CPU and 32GB memory when comparing their running times.

Validation with random forest model on scRNA-seq reference data.

For each cell type, top 100 marker genes are kept as the features to train the random forest model, resulting in 1118 total features whose expression values are then normalized into counts per thousand. 1000 trees have been used to train the model, while other parameters are as the default in 'randomForest' package. To evaluate the similarity between decomposition methods and the prediction of the random forest model, we apply bootstrapping where 100 spots in the ST data are randomly chosen for 100 times. The multi-class AUROC are calculated based on their results in each part by the 'multiclass.roc' function in 'pheatmap' package. The overall accuracy of the results is also calculated based on whether a primary component is predicted correctly by the random forest model. The same procedure is rerun on the top 10% of spots whose primary components take the largest part. The negative control group is created by randomly generating the prediction scores of each cell type from the uniform distribution $U(0,1)$. The values are then scaled to make sure the sum of the predictions in one spot equals to 1.

Co-kriging models spatial cell-type compositional data.

We select top 10 cell types which have the highest prevalence among all spots in the compositional data generated by RCTD full mode. 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 matrix of log-ratio variograms is then calculated with 15 distance bins. Next, A compositional linear model of coregionalization with a nugget effect term and 3 Gaussian variogram terms whose respective effective as 600, 1300, and 3000 micrometers, is trained based on the data and the matrix. The parameters are determined by minimizing the residual of the model.

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

501 the ground distances, and the proportions of each cell-type composition are inputted as the weights when
502 comparing to a gene's spatial distribution. R package 'emdist' is used to calculate EMD in the simulations.
503
504 **Spatial ST data simulation.**
505 Simulation data are generated based on the following steps: 1) A 20x20 grids is created and its distance
506 matrix is calculated; 2) 5 clusters of overlapping marker genes, 20 in each cluster, are created, and their
507 relationship between the number of corresponding 'cell type' c_i is then generated with a negative binomial
508 model:

509 $$c_{i,k} \sim \mathrm{NB}(\sum_g a_{i,g} y_{g,k}, \sum_g a_{i,g} y_{g,k})$$

510 where $y_{g,k}$ is the scaled expression level of gene g at spot k. $c_{i,k}$ is the number of the type i cells in spot k.
511 $a_{i,g}$ are the correlation parameters between gene g and cell type i: if the gene is the marker of the given
512 cell type, $a_{i,g} \sim \mathrm{U}(500, 1000)$; otherwise, $a_{i,g'} \sim \mathrm{U}(-50, 50)$; 3) the marker genes' spatial distributions
513 are generated based on the bivariate normal distribution. The centers of 5 clusters are randomly simulated
514 $m_i \sim \mathrm{U}^2(0, 20)$ such that they are at least 5 units away from each other. The centers of genes' spatial
515 distribution follow $\mu_{i,g} \sim m_i + \mathrm{N}(0, 1)$, and their variance $\Sigma \sim \Sigma_0 + \mathrm{N}(0, 0.5)$. 4) Another 5 clusters of
516 surrounding genes are also generated in order to compare the performance on detecting paracrine regulators.
517 For cell type i, its corresponding surrounding genes have the center m_i and their expression levels
518 $y'_{g,k}$ follow a normal distribution $\mathrm{N}(R_{g,k} + \mathrm{N}(0,1), 2)$ where $R_{g,k}$ is the distance parameter that controls
519 the distance between paracrine signals and target cell types. The background expression level in the rest
520 spots $y_0 \sim \mathrm{U}(0.1, 0.5)$. 5) as the negative control group, 30 universal genes are generated whose spatial
521 expression patterns follow $\mathrm{U}(0, 1) + 0.5$. 6) To generate the cell-type compositional data, we divide the
522 cell count matrix by the total number of cells in each spot after halving the values which are less than 40.
523 In total, 50 simulations have been run where the EMD, Pearson and Spearman correlation, and SPER scores
524 are respectively calculated. The resulting matrices are used to predict the spatial dependence between genes

525 and cell types. For each type, the ROC curve is plotted in distinguishing its markers and paracrine signals
526 from the other genes.
527
528 **Spatial-dependent signals detection**.
529 The SPER score matrix, Pearson correlation matrix and Spearman correlation matrix are used to filter the
530 candidate genes and spatial-dependent signals with the help of scRNA-seq data as the reference of
531 expression levels. Only the intersect of genes detected by both ST data and scRNA-seq data and with total
532 UMI counts greater than 40 in the ST data are kept, resulting in a number of 13436 genes. When studying
533 the enrichment of extracellular and ligand genes, we used the hypergeometric test to see if they are enriched
534 within the gene set given thresholds. In both cases, the population size is 13436; the number of successes
535 in the population is 1784 for the extracellular and 1888 for ligands. The sample size and the number of
536 successes in the sample are dependent on the expression-level and measurement-level thresholds. If the p-
537 values calculated from the tests are less than 0.05, we will take it as significantly enriched.
538
539 **Calculation of SPER scores**
540 Two key factors that affect SPER score calculation are the choices of distance bin width and their
541 corresponding weights. In the default setting, we use the average distance between two spots as the width
542 of bins in accordance with the technical parameters, which is the smallest meaningful distance width in the
543 Visium data. This choice is certainly adjustable if one wants to discover the signals within a larger distance
544 scale. However, the increase in the distance bin width will reduce SPER's ability to capture the short-range
545 signals, so a trade-off is necessary when making the decision.
546 The weight parameters on each distance group are also essential when integrating the SPER scores. One
547 can modify the weight on each group in order to detect different types of signals: for example, in a situation
548 where a 1mm-range signal is what we are looking for, we can then weigh more on the distance group around
549 1000μm. To notice, the change of weights will generally affect the SPER score performance by highlighting
550 the signals fitted the weight well and underrating the potential signals that function at a distance away from

551 the high-weight groups. Particularly, when all weights lie at the zero-distance group, SPER will just
552 function as a metric of spatial co-localization between genes and cell types. In practice, users can choose
553 more than one set of weights to integrate SPER scores to catch these different signals.
554

555 **Acknowledgements**
556 The authors thank Leslie Gaffney for preparing and generation of figures for this article, Evan Lemire for
557 programming assistance, and Yu Sun for helping improve the data visualization.
558

559 **Data and Code availability**
560 An R package to implement the SPER algorithm is available for download at
561 github.com/TianxiaoNYU/SPER. Extracellular gene set in mouse is available from the COMPARTMENTS
562 database: https://compartments.jensenlab.org/Downloads; scRNA-seq reference data is obtained from
563 Allen Institute for Brain Science: https://portal.brain-map.org/atlases-and-data/rnaseq; Visium spatial
564 transcriptomics data from mouse brain is available from 10X Genomics dataset:
565 https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-
566 standard-1-0-0. An R markdown tutorial to help generate the main results of this article using SPER package,
567 as well as all other necessary data files in this study can be found from the below Dropbox link:
568 https://www.dropbox.com/scl/fo/pep07jjid71rr72voeeik/h?rlkey=2agyojo1ugf5l61bqtal6hjy0&dl=0.
569

570 **References**

[1] Johnson, T.S., Xiang, S., Dong, T., "Combinatorial analyses reveal cellular composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer's Disease," *Scientific Reports,* vol. 11, p. 353, 2021.

[2] Knoop, K.A., "Goblet cells: multifaceted players in immunity at mucosal surfaces," *Mucosal Immunology,* no. 11, p. 1551–1557, 2018.

[3] Carr, E., Dooley, J., Garcia-Perez, J. , "The cellular composition of the human immune system is shaped by age and cohabitation," *Nature Immunology,* vol. 17, p. 461–468, 2016.

[4] Ansell, S.M., Vonderheide, R.H., "Cellular composition of the tumor microenvironment," *American Society of Clinical Oncology Educational Book,* vol. 33, 2013.

[5] Hewitt, R., Lloyd, C., "Regulation of immune responses by the airway epithelial cell landscape.," *Nature Reviews Immunology,* vol. 21, no. 6, pp. 347-362, 2021.

[6] M. Wills-Karp, "Interleukin-13: central mediator of allergic asthma," *Science,* vol. 282, no. 5397, pp. 2258-61, 1998.

[7] Palla, G., Fischer, D.S., Regev, A., "Spatial components of molecular tissue biology," *Nature Biotechnology,* no. 40, p. 308–318, 2022.

[8] C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo and A. Haque, "An introduction to spatial transcriptomics for biomedical research," *Genome Medicine,* no. 14, p. 68, 2022.

[9] 10X Genomics, "Mouse Brain Serial Section 1 (Sagittal-Anterior)," 2 12 2019. [Online]. Available: https://www.10xgenomics.com/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-0-0. [Accessed 10 9 2021].

[10] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen and E. Z. Macosko, "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution," *Science,* vol. 363, pp. 1463-1467, 2019.

[11] E. A. G. Baker, "Power analysis for spatial omics," *bioRxiv,* 2022.

[12] Dries, R., Zhu, Q., Dong, R., "Giotto: a toolbox for integrative analysis and visualization of spatial expression data," *Genome Biology,* vol. 22, p. 78, 2021.

[13] R. Tolosana-Delgado, "Geostatistics for Compositional Data: An Overview," *Mathematical Geosciences,* vol. 51, pp. 485-526, 2019.

[14] D. G. Krige, "A statistical approach to some basic mine valuation problems on the Witwatersrand," *Journal of the Southern African Institute of Mining and Metallurgy,* vol. 52, p. 6, 1951.

[15] Stuart T., et al., "Comprehensive Integration of Single-Cell Data," *Cell,* vol. 177, no. 7, pp. 1888-1902.e21, 2019.

[16] Dong, R., Yuan, GC., "SpatialDWLS: accurate deconvolution of spatial transcriptomic data," *Genome Biology,* vol. 21, p. 145, 2021.

[17] Cable, D.M., Murray, E., Zou, L.S., "Robust decomposition of cell type mixtures in spatial transcriptomics," *Nature Biotechnology,* vol. 40, p. 517–526, 2022.

[18] Harvey, S.A., Smith, J.C., "Visualisation and Quantification of Morphogen Gradient Formation in the Zebrafish," *PLOS Biology,* vol. 7, no. 5, p. e1000101, 2009.

[19] E.M. Lifshitz, L.P. Pitaevski, "The Diffusion Approcimation," in *Physical Kinetics*, Oxford, Butterworth-Heinemann, 1981, pp. 89-114.

[20] D. E. Myers, "Matrix formulation of co-kriging," *Journal of the International for Mathematical Geology,* pp. 249-257, 1982.

[21] Y. Rubner, Tomasi and Carlo, "Metric for distributions with applocation to image databases," in *Sixth international conference on computer vision*, 1998.

[22] Finak, G., et al., "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biololgy,* vol. 16, p. 278, 2015.

[23] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee and K. Smith, "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.," *Nature Neuroscience,* vol. 19, no. 2, pp. 335-346, 2016.

[24] Binder, JX., et al., "COMPARTMENTS: unification and visualization of protein subcellular localization evidence," *Database (Oxford),* vol. 2014, p. bau012, 2014.

[25] Ramilowski, J., et al., "A draft network of ligand–receptor-mediated multicellular signalling in human," *Nature Communications,* vol. 6, p. 7866, 2015.

[26] Efremova, M., et al., "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes," *Nature Protocols,* vol. 15, p. 1484–1506, 2020.

[27] S. Frazer, J. Prados, M. Niquille, C. Cadihac, F. Markopoulos, L. Gomez, U. Tomasello, L. Telley, A. Holtmaat, D. Jabaudon and A. Dayer, "Transcriptomic and anatomic parcellation of 5-HT3AR expressing cortical interneuron subtypes revealed by single-cell RNA sequencing," *Nature Communications,* vol. 8, no. 1, p. 14219, 2017.

[28] A. Ito-Ishida, S. Okabe and M. Yuzaki, "The role of Cbln1 on Purkinje cell synapse formation," *Neuroscience research,* pp. 64-68, 2014.

[29] P. Han, Y. She, Z. Yang, M. Zhunag, Q. Wang, X. Luo, C. Yin, J. Zhu, S. R. Jaffrey and S.-J. Ji, "Cbln1 regulates axon growth and guidance in multiple neural regions," *PLOS Biology,* p. e3001853, 2022.

[30] N. Barker, J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters and H. Clevers, "Identification of stem cells in small intestine and colon by marker gene Lgr5," *Nature,* vol. 449, no. 7165, pp. 1003-1007, 2007.

[31] Sungwoo Bae, Hongyoon Choi, Young Tae Kim, "CellDART: cell type inference by domain adaptation of single-cell and spatial transcriptomic data," *Nucleic Acids Research,* 2022.

[32] Lopez, R., Li, B., Keren-Shaul, H. , "DestVI identifies continuums of cell types in spatial transcriptomics data," *Nature Biotechnology,* 2022.

[33] Lee, Y., et al., "XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment," *Science Advances,* vol. 7, p. eabg4755, 2021.

[34] M. Zhang, S. W. Eichhorn, B. Zingg, Z. Yao, K. Cotter, H. Zeng, H. Dong and X. Zhuang, "Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH," *Nature,* vol. 598, no. 7879, pp. 137-143, 2021.

[35] V. Petuhov, R. J. Xu, R. A. Soldatov, P. Cadinu, K. Khodosevich, J. R. Moffitt and P. V. Kharchenko, "Cell segmentation in imaging-based spatial transcriptomics," *Nature Biotechnology,* vol. 40, no. 3, pp. 345-354, 2022.

[36] Ordoñez CL, "Mild and moderate asthma is associated with airway goblet cell hyperplasia and abnormalities in mucin gene expression," *Am J Respir Crit Care Med,* vol. 163, no. 2, pp. 517-23, 2001.

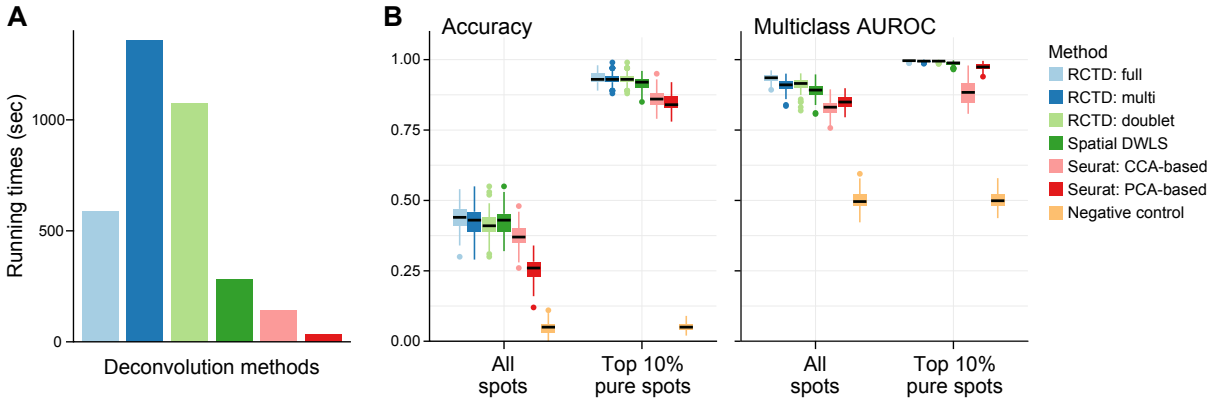
[37] S. He, H. Zhou, X. Zhu, S. Hu, M. Fei, D. Wan, W. Gu, X. Yang, D. Shi, J. Zhou, J. Zhou, Z. Zhu, L. Wang, D. Li and Y. Zhang, "Expression of Lgr5, a marker of intestinal stem cells, in colorectal cancer and its clinicopathological significance," *Biomedicine & pharmacotherapy,* p. 507–513, 2014.

[38] L. Yang, H. Tang, Y. Kong, X. XIe, J. CHen, C. Song, X. Liu, F. Ye, N. Li, N. Wang and X. Xie, "LGR5 Promotes Breast Cancer Progression and Maintains Stem-Like Cells Through Activation of Wnt/β-Catenin Signaling," *Stem Cells,* vol. 33, no. 10, pp. 2913-2924, 2015.

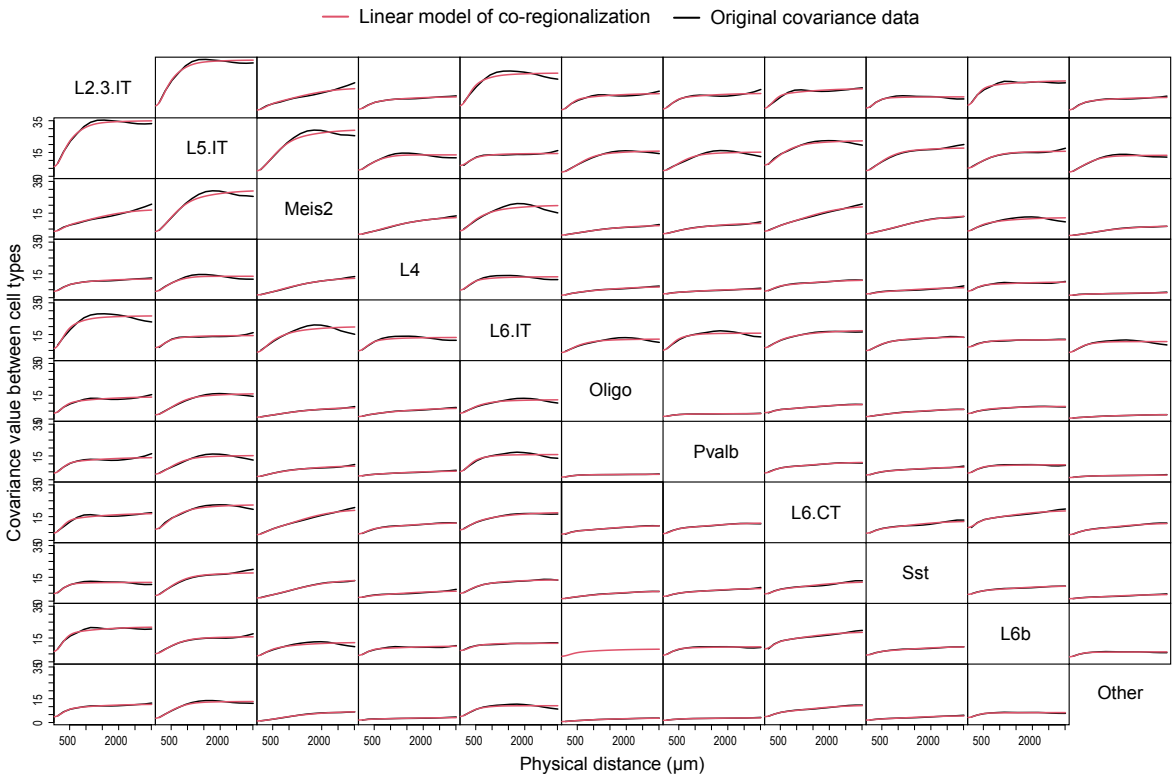
571

572

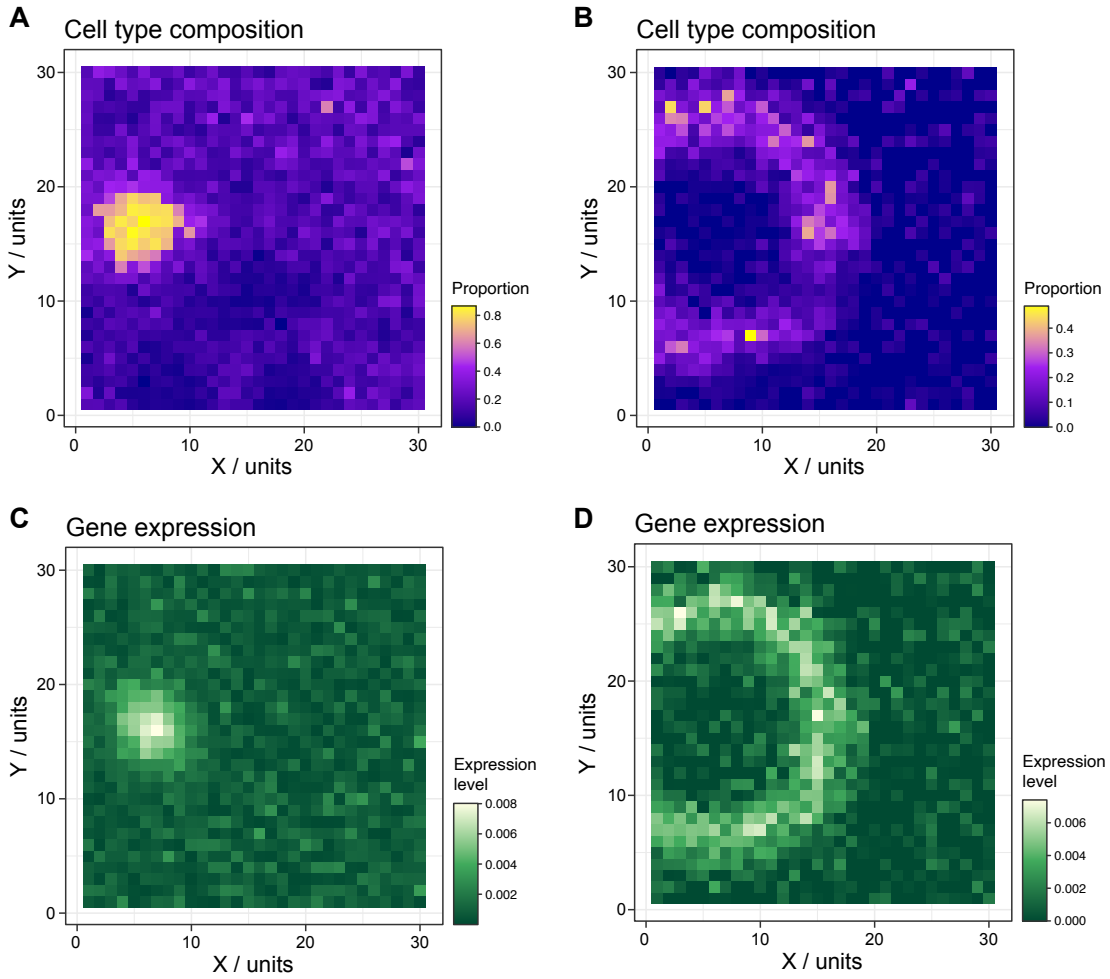
SUPPLEMENTARY FIGURES:



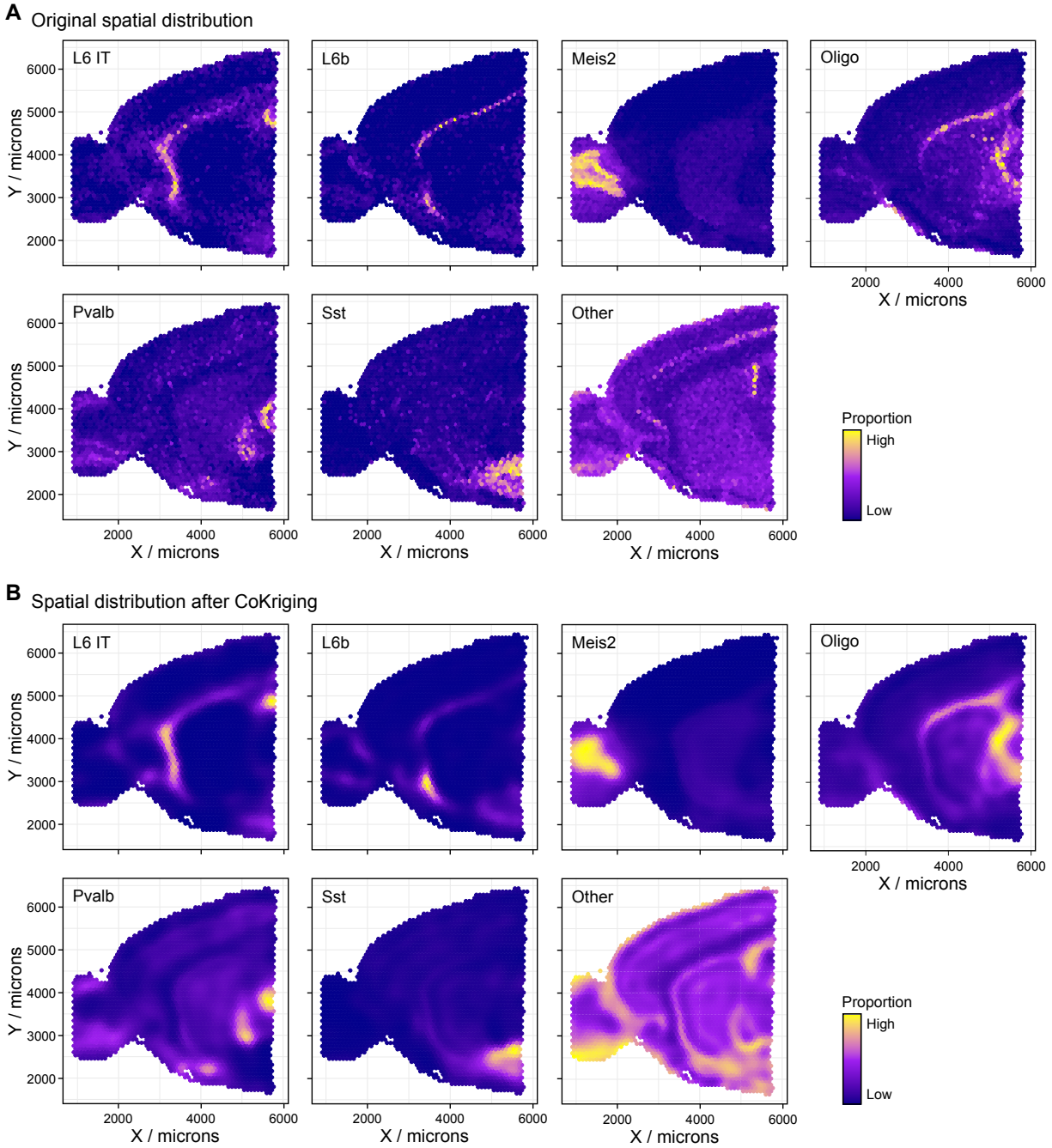
Supplementary Figure 1. Comparison of published spatial cell type deconvolution methods. (A) The runtimes in seconds of each cell type deconvolution methods (color legend). Technical details are described in Methods. (B) The overall accuracy and multiclass AUROC on the primary component prediction. The negative control group is marked as orange.



581 **Supplementary Figure 2. Co-kriging fits visualized as variograms**. Co-kriging method fits on the 10
582 neuronal cell types from an example dataset of compositional data from mouse brain. A compositional
583 linear model of co-regionalization (red line) with a 'nugget' effect term for short scale randomness and 3
584 Gaussian variogram terms to model longer term effects (600, 1300, and 3000 μm) showed excellent fit to
585 the original covariance data (black line) collected and calculated at 15 uniformly distributed lags. The
586 labels at the diagonal identify the cell type for each row and column, while 'Other' represents the sum of
587 all remaining cell types. The x-axis in each subpanel represents the physical distance in μm, and y-axis
588 represents the relative value of the covariance between each pair of cell types.
589



Supplementary Figure 3. Examples of simulated cell-types and signaling genes distributions. (A, B) Two examples of the simulated cell-type spatial compositional data. Color (purple-yellow) represents the proportion of the given cell types, where cell A follows a 2-D gaussian distribution with extra background noise and cell B follows a circular-designed gaussian distribution (details in Methods). (C, D) Two examples of the simulated signaling genes' spatial distribution. Color (green) represents the expression level of the gene C and D (details in Methods). Specifically, gene C is cell-A-specific, and gene D is cell-B-specific. Mentioned in the Results, we proposed two types of signaling genes: 'overlap' (cell A and gene C), and 'paracrine' (cell A and gene D). In each subplot, X and Y coordinates represent the physical location in the spatial simulation.

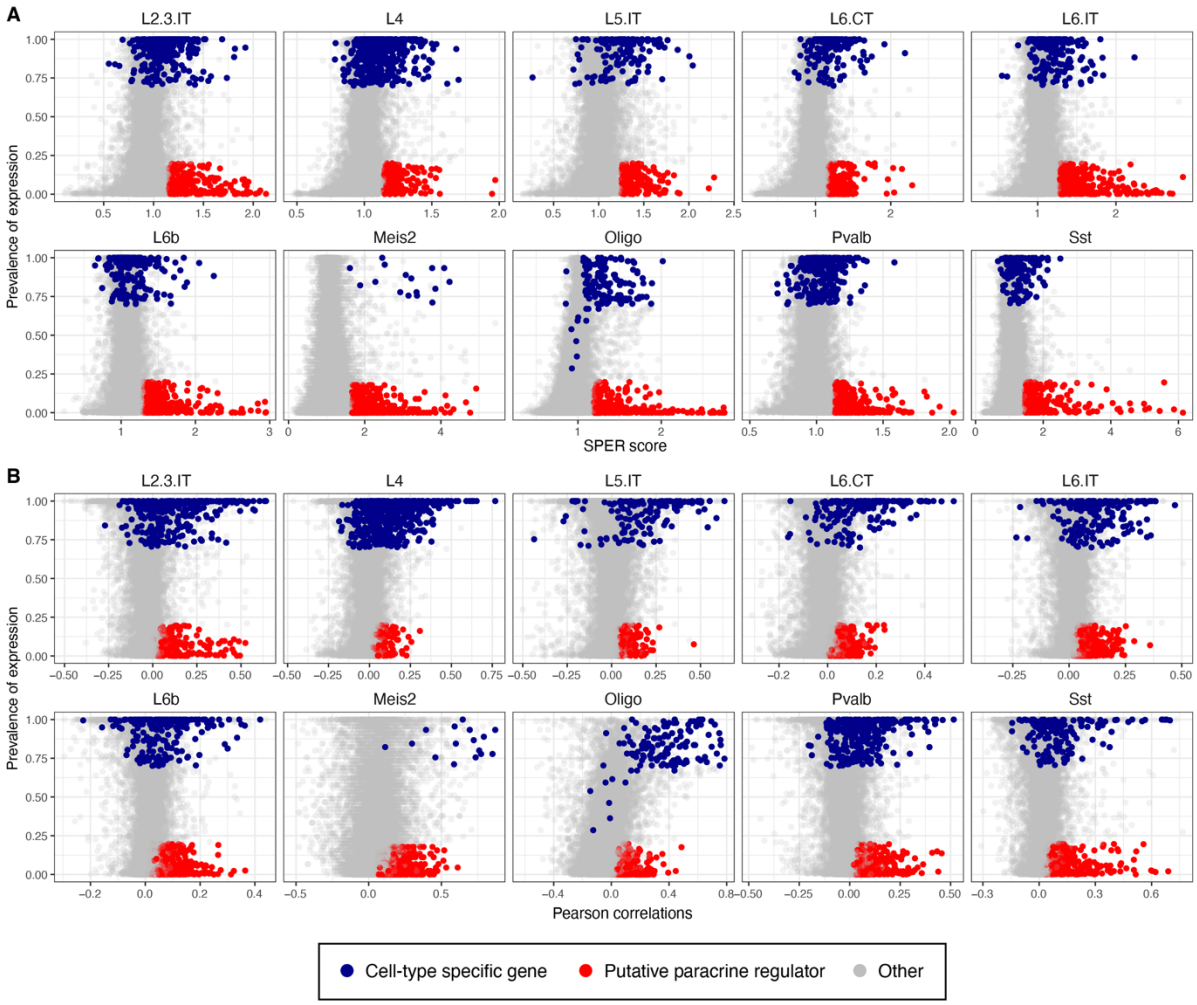


599 **Supplementary Figure 4. Co-kriging models effectively denoise spatial compositional data**. Spatial
600 CoDa of the other primary cell types (in addition to those shown in **Fig. 4**) as well as the 'Other'
601 decomposed by RCTD full mode before and after Co-kriging modeling. The top 10 most abundance cell
602 types are modeled individually and the other X cell types are summed as the 'Other' category. In each

603 subplot, X and Y-axis represent the physical coordinates of spots in μm, and the color represents the

604 compositional proportion from 0 to 1. The subtitles label the cell type.

605



606

607 **Supplementary Figure 5. Comparison of trivial spatial associations and genuine putative regulators**
608 **between SPER and Pearson correlation.** Scatter plots show the distribution of SPER (**A**) and Pearson
609 correlation (**B**) scores (x axis) with the proportion of cells of each type (plots) in which each gene (points)
610 is detected (y axis) in scRNA-seq data. The dark blue dots mark the 'trivial associations': marker genes for
611 each cell type (found by MAST, details in Methods), these are trivial because the expression level of a cell's
612 marker gene is trivially associated with that cell's composition, since that cell is a source of that transcript.
613 Red dots on the other hand represent the candidate signal genes whose expression prevalence is less than
614 20% in their target cell types and measurement scores larger than 95% quantiles, and thus represent non-
615 trivial spatial associations.

616