Comprehensive visualization of cell-cell interactions in single-cell and spatial transcriptomics with NICHES

Authors

Micha Sam Brickman Raredon[1,2,†,*] and Junchen Yang[3,†], Neeharika Kothapalli[4], Naftali Kaminski[4], Laura E. Niklason[1,5], Yuval Kluger[3,6,7,*]

Affiliations

[1] Department of Biomedical Engineering, Yale University, New Haven, CT, 06511
[2] Medical Scientist Training Program, Yale School of Medicine, New Haven, CT, 06511
[3] Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, 06511
[4] Pulmonary, Critical Care, and Sleep Medicine, Yale School of Medicine, New Haven, CT 06511
[5] Department of Anesthesiology, Yale School of Medicine, New Haven, CT, 06511
[6] Department of Pathology, Yale School of Medicine, New Haven, CT, 06511
[7] Applied Mathematics Program, Yale University, New Haven, CT, 06511
[†] These authors contributed equally
[*] Corresponding author

Abstract

Summary: Recent years have seen the release of several toolsets that reveal cell-cell interactions from single-cell data. However, all existing approaches leverage mean celltype gene expression values, and do not preserve the single-cell fidelity of the original data. Here, we present **NICHES** (**N**iche **I**nteractions and **C**ommunication **H**eterogeneity in **E**xtracellular **S**ignaling), a tool to explore extracellular signaling at the truly single-cell level. NICHES allows embedding of ligand-receptor signal proxies to visualize heterogeneous signaling archetypes within cell clusters, between cell clusters, and across experimental conditions. When applied to spatial transcriptomic data, NICHES can be used to reflect local cellular microenvironment. NICHES can operate with any list of ligand-receptor signaling mechanisms and is compatible with existing single-cell packages and pseudotime techniques. NICHES is also a user friendly and extensible program, allowing rapid analysis of cell-cell signaling at single-cell resolution.
Availability and implementation: NICHES is an open-source software implemented in R under academic free license v3.0 and it is available at github.com/msraredon/NICHES.
Contact: michasam.raredon@yale.edu; yuval.kluger@yale.edu

1. Background

Cellular phenotype across tissues and organs is heavily influenced by the biological microenvironment in which a given cell resides (Baccin, et al., 2020; Davidson, et al., 2020; McCarthy, et al., 2020; Nabhan, et al., 2018; Qadir, et al., 2020; Rodda, et al., 2018; Tikhonova, et al., 2020; Zhou, et al., 2018). Understanding the influence of cell-cell signaling on cell phenotype is a major goal in developmental and tissue biology and has profound implications for our ability to engineer tissues and next-generation cellular therapeutics. Single-cell technologies, which capture information both from individual cells and their surrounding cellular environment at the same time, are uniquely suited to exploring phenotype-environment relations. Many techniques are available to extract and prioritize extracellular signaling patterns from single-cell data, including CellPhoneDB, NicheNet, CellChat, Connectome, SingleCellSignalR, iTALK, iCELLNET, Cellinker, CellCall, and PyMINEr, among others (Browaeys, et al., 2019; Cabello-Aguilar, et al., 2020; Efremova, et al., 2020; Jin, et al., 2020; Noël, et al., 2020; Raredon, et al., 2021; Tyler, et al., 2019; Wang, et al., 2019; Zhang, et al., 2021; Zhang, et al., 2021). However, current computational methods are based on mean expression values taken from each celltype cluster. Mean expression representation does not take full advantage of the single-cell resolution of the original measurements, thereby obscuring the rich repertoire of signaling patterns between cells. The field can benefit from a tool that assesses cell-cell signaling at the *truly single-cell level*, so that intra- and inter-cluster signaling patterns can be inferred from the observed data.

Here, we describe NICHES (**N**iche **I**nteractions and **C**ommunication **H**eterogeneity in **E**xtracellular **S**ignaling), a computational workflow to characterize cellular interactions in ligand-receptor signaling-space at the single-cell level. NICHES is designed for analysis of two types of cellular interactions: cell-cell signaling (defined as the signals passed between cells, determined by the product of the ligand expression of the sending cell and the receptor expression of the receiving cell) and cellular niche (defined as the signaling input to a cell, determined by the sum of the ligand profiles of the surrounding cells and the receptor profile of the receiving cell). NICHES allows embedding, exploration, and analysis of these interactions using already-existing single-cell software, such as Seurat and Scanpy (Butler, et al., 2018; Wolf, et al., 2018).

2. Approach

NICHES takes single-cell data as input and constructs matrices where the rows are extracellular ligand-receptor signaling mechanisms and the columns are cell-cell extracellular signaling interactions (Fig 1A-C). Cell-cell interactions are represented as columns whose entries are calculated by multiplying ligand expression on the sending cell with receptor expression on the receiving cell, for each mechanism (Fig 1B). Cellular microenvironment, or the niche of each cell, is represented as columns that are calculated by summing all incoming cell-cell edges which land on the respective receiving cell (Fig 1C). Row names are based on a ground-truth ligand-receptor mechanism list which can be customized by the user. NICHES provides built-in access to ligand-receptor lists from the OmniPath and FANTOM5 databases (Ramilowski, et al., 2015; Türei, et al., 2021)and is compatible with custom mechanism lists containing any number of ligand or receptor subunits.

When applied to spatial transcriptomic data, interactions may be constrained to those occurring between spatial neighbors. When applied to single-cell data, NICHES samples unique cell pairs from each celltype-celltype cross. Detailed methods and customizable options are provided in Supplemental Information.

3. Application

3.1 Advantages of NICHES over Existing Techniques

NICHES facilitates discovery of intra-cluster signaling heterogeneity and subpopulations which can be hidden by existing cell-cell signaling inference tools (Supp Fig 1A-C, Supplemental Text).

Because NICHES does not leverage cluster-wise mean values, significant differences between experimental conditions may be detected when expression distribution shifts but the mean value remains similar (Supp Fig 1E-I, Supplemental Text). NICHES also allows users to visualize changes in niche signaling due to the addition or loss of a celltype, a task which is not possible with current methods (Supp Fig 1J-L, Supplemental Text).

3.2 Cell-Cell Signaling Atlases

NICHES allows comprehensive visualization of ligand-receptor patterns that are present in single-cell data (Fig 1D-F). A uniform sample is taken of every celltype-celltype interaction resulting in a cell-cell signaling atlas that can be viewed via low-dimensional embedding (Fig 1D). Each celltype-celltype interaction displays a specific signaling signature with some degree of intra-relationship heterogeneity (Fig 1E). A given celltype-celltype cross may then be subclustered (Fig 1F) to further explore heterogeneous relationships and identify mechanisms specific to subtypes of cell crosses (Tgfb1-Cav1 is specific to Subcluster 2 in this instance.)

3.3 Mapping Local Microenvironment in Spatial Atlases

NICHES quantifies local microenvironments in spatial transcriptomic datasets. Interactions may be limited to spatial nearest neighbors, allowing an estimation of local niche for each transcriptomic spot to be visualized in low-dimensional space (Fig 1G). Each celltype displays a stereotyped niche signature with observable intra-cluster heterogeneity (Fig 1H). Our approach provides a broad picture of cell signaling, while identifying tightly localized celltype-specific niche interactions (Fgf1-Fgfr2 in this instance is found to be specific to the oligodendrocyte microenvironment, in agreement with existing literature (Furusho, et al., 2020; Furusho, et al., 2015).) Niche interactions of interest may be visualized in spatial context (Fig 1I). Sub-clustering may be performed to identify intra-celltype microenvironment heterogeneity correlated with tissue boundaries and transition regions (Supp Fig 2).

3.4 Niche Signaling Changes During Differentiation

NICHES may be applied to time-course data to explore how niche signaling changes in pseudotime and over differential branching trajectories (Fig 1J,K). The niche for each cell is calculated by summing all incoming cell-cell interactions from that cell's respective batch or timepoint, across each signaling mechanism queried, for a given receiving cell. Pseudotemporal ordering then allows graphing of extracellular niche signals that are associated with lineage branches and differentiation trajectories as a celltype develops within context (Fig 1L).

4. Conclusion

NICHES is a simple but powerful approach to explore cell-cell signaling interactions in single-cell and spatial transcriptomic data. NICHES supplements the capabilities of current techniques, allowing single-cell resolution of niche signaling and cell-cell interactions, thereby establishing rich representations to analyze environment-phenotype relationships in tissues.

Figures

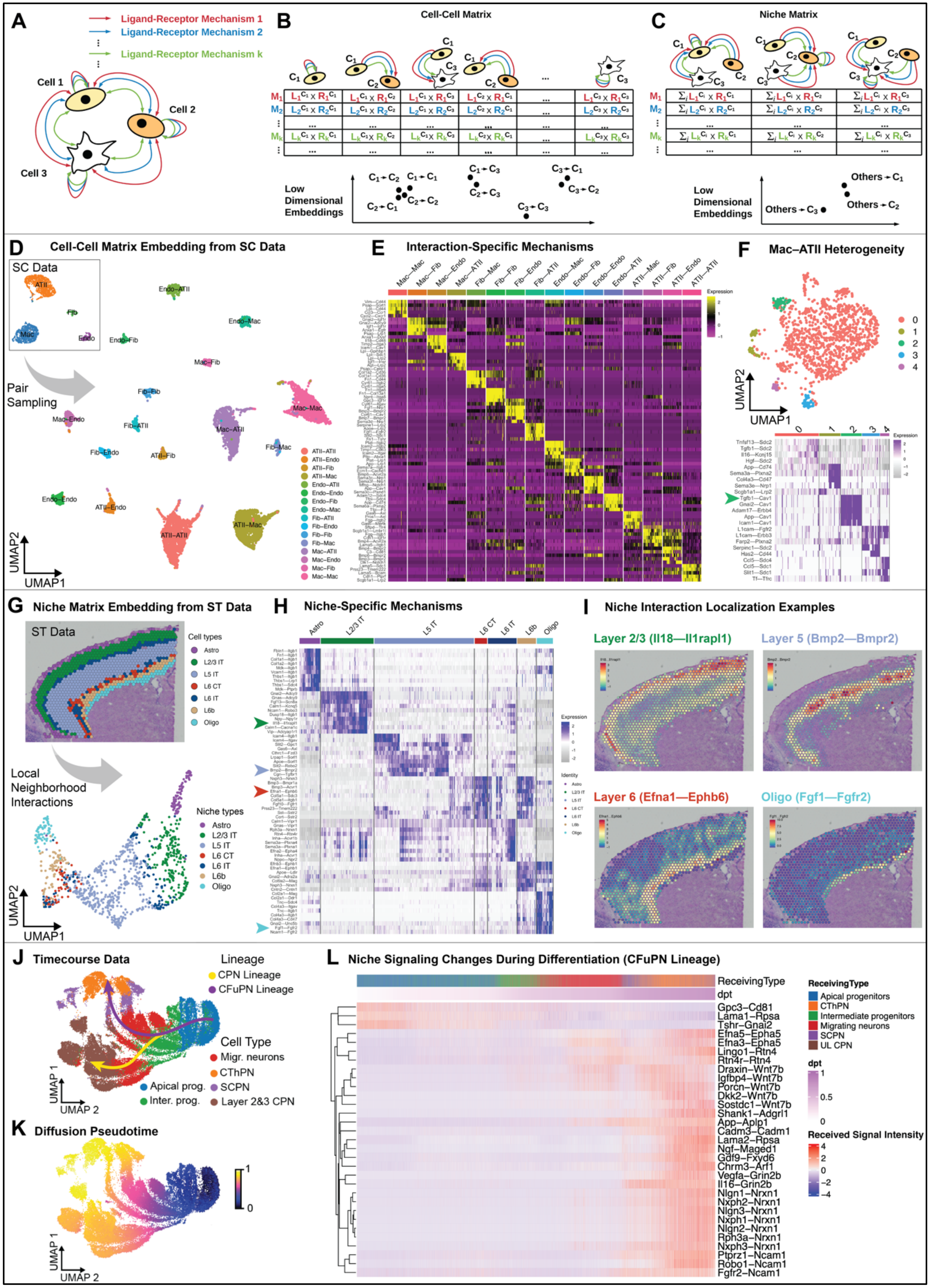
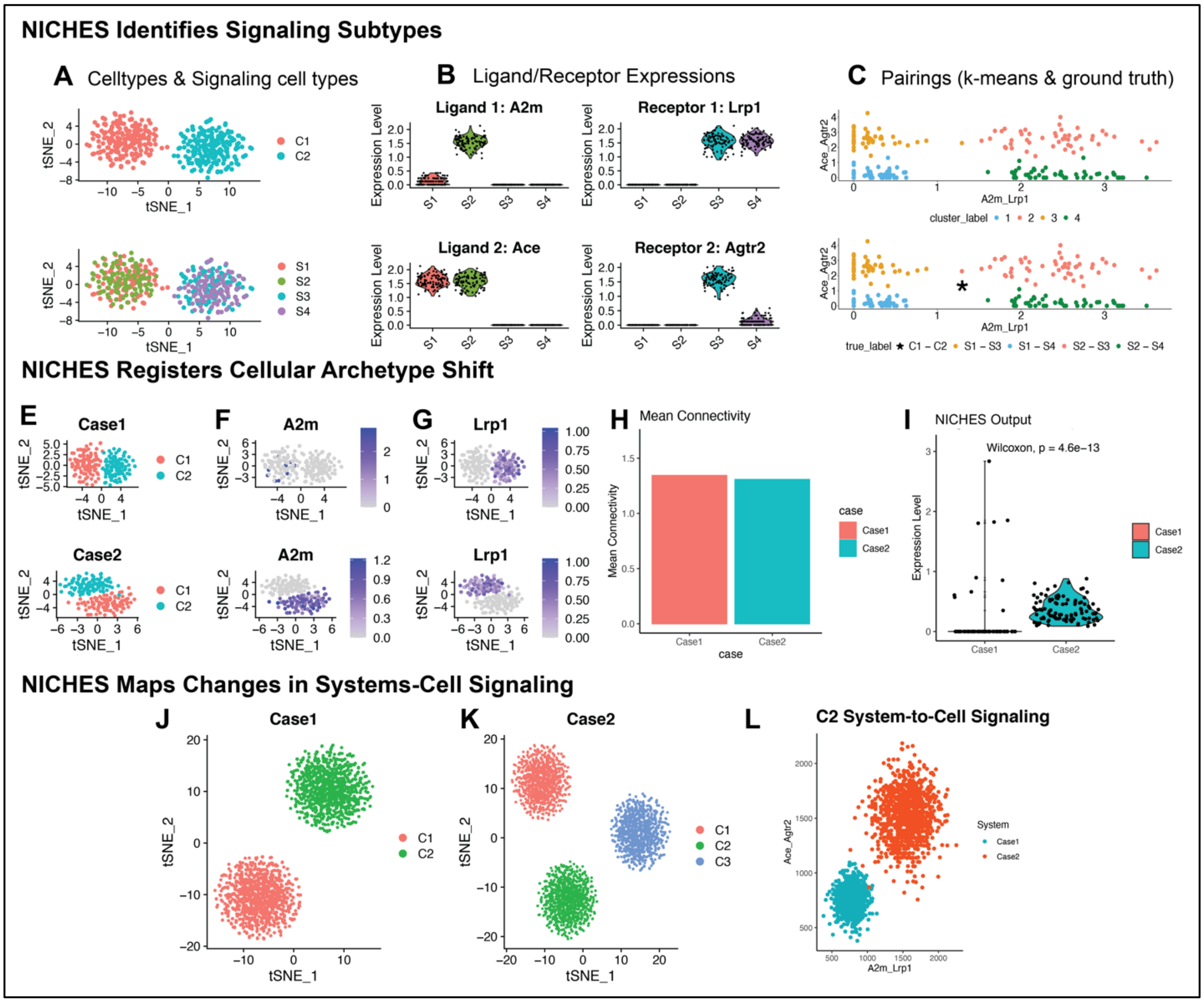

A
Ligand-Receptor Mechanism 1
Ligand-Receptor Mechanism 2
Ligand-Receptor Mechanism k
Cell 1
Cell 2
Cell 3
B
Cell-Cell Matrix
M1
M2
Mk
Low Dimensional Embeddings
C1→C2
C1→C1
C1→C3
C3→C1
C2→C2
C2→C1
C2→C3
C3→C2
C3→C3
C
Niche Matrix
Others→C1
Others→C3
Others→C2
D
Cell-Cell Matrix Embedding from SC Data
SC Data
ATII
Fib
Mac
Endo
Pair Sampling
Endo–ATII
Endo–Mac
Endo–Fib
Mac–Fib
Fib–Fib
Mac–Mac
Mac–Endo
Fib–ATII
Mac–ATII
Fib–Mac
Fib–Endo
ATII–Fib
Endo–Endo
ATII–Endo
ATII–ATII
ATII–Mac
ATII–ATII
ATII–Endo
ATII–Fib
ATII–Mac
Endo–ATII
Endo–Endo
Endo–Fib
Endo–Mac
Fib–ATII
Fib–Endo
Fib–Fib
Fib–Mac
Mac–ATII
Mac–Endo
Mac–Fib
Mac–Mac
UMAP2
UMAP1
E
Interaction-Specific Mechanisms
Mac–Mac
Mac–Fib
Mac–Endo
Mac–ATII
Fib–Mac
Fib–Fib
Fib–Endo
Fib–ATII
Endo–Mac
Endo–Fib
Endo–Endo
Endo–ATII
ATII–Mac
ATII–Fib
ATII–Endo
ATII–ATII
Expression
F
Mac–ATII Heterogeneity
0
1
2
3
4
UMAP2
UMAP1
Expression
Tnfsf13—Sdc2
Tgfb1—Sdc2
Il16—Kcnj15
Hgf—Sdc2
App—Cd74
Sema3a—Plxna2
Col4a3—Cd47
Sema3e—Nrp1
Scgb1a1—Lrp2
Tgfb1—Cav1
Gnai2—Cav1
Adam17—Erbb4
App—Cav1
Icam1—Cav1
L1cam—Fgfr2
L1cam—Erbb3
Farp2—Plxna2
Serpinc1—Sdc2
Has2—Cd44
Ccl5—Sdc4
Ccl5—Sdc1
Slit1—Sdc1
Tf—Tfrc
G
Niche Matrix Embedding from ST Data
ST Data
Cell types
Astro
L2/3 IT
L5 IT
L6 CT
L6 IT
L6b
Oligo
Local Neighborhood Interactions
Niche types
Astro
L2/3 IT
L5 IT
L6 CT
L6 IT
L6b
Oligo
UMAP2
UMAP1
H
Niche-Specific Mechanisms
Astro
L2/3 IT
L5 IT
L6 CT
L6 IT
L6b
Oligo
Expression
Identity
I
Niche Interaction Localization Examples
Layer 2/3 (Il18—Il1rapl1)
Layer 5 (Bmp2—Bmpr2)
Layer 6 (Efna1—Ephb6)
Oligo (Fgf1—Fgfr2)
J
Timecourse Data
Lineage
CPN Lineage
CFuPN Lineage
Cell Type
Migr. neurons
CThPN
Apical prog.
SCPN
Inter. prog.
Layer 2&3 CPN
UMAP 1
UMAP 2
K
Diffusion Pseudotime
1
0
UMAP 1
UMAP 2
L
Niche Signaling Changes During Differentiation (CFuPN Lineage)
ReceivingType
dpt
Gpc3—Cd81
Lama1—Rpsa
Tshr—Gnai2
Efna5—Epha5
Efna3—Epha5
Lingo1—Rtn4
Rtn4r—Rtn4
Draxin—Wnt7b
Igfbp4—Wnt7b
Porcn—Wnt7b
Dkk2—Wnt7b
Sostdc1—Wnt7b
Shank1—Adgrl1
App—Aplp1
Cadm3—Cadm1
Lama2—Rpsa
Ngf—Maged1
Gdf9—Fxyd6
Chrm3—Arf1
Vegfa—Grin2b
Il16—Grin2b
Nlgn1—Nrxn1
Nxph2—Nrxn1
Nlgn3—Nrxn1
Nxph1—Nrxn1
Nlgn2—Nrxn1
Rph3a—Nrxn1
Nxph3—Nrxn1
Ptprz1—Ncam1
Robo1—Ncam1
Fgfr2—Ncam1
ReceivingType
Apical progenitors
CThPN
Intermediate progenitors
Migrating neurons
SCPN
UL CPN
dpt
1
0.5
0
Received Signal Intensity
4
2
0
−2
−4

Figure 1: NICHES allows analysis of cell-cell interactions with single-cell resolution. A) A set of cells may interact through many different ligand-receptor mechanisms. B) NICHES represents cell-cell interactions as columns whose entries are calculated by multiplying ligand expression on the sending cell with receptor expression on the receiving cell, for each signaling mechanism. Low-dimensional embeddings may then be made of cell-cell interactions. Note schematic clustering of similar profiles. C) Cellular microenvironments, or niches, of each cell are represented as columns calculated by summing all incoming cell-cell columns landing on the receiving cell. This allows low-dimensional embedding of a proxy for sensed microenvironment for each cell. D) NICHES analysis of single-cell (SC) data of four cell types co-localized in the rat pulmonary alveolus yields a quantitative cell-cell signaling atlas visualized by low-dimensional embedding. E) Biologically relevant signaling markers can be identified specific to each celltype-celltype interaction. Because single-cell fidelity is preserved, NICHES allows observation of fine intra- and inter-cluster heterogeneity unobservable using mean-wise techniques. F) Subclustering of a single celltype-celltype cross allows observation of fine heterogeneity and identification of mechanisms employed by only a specific subset of cell pairings (Tgfb1-Cav1, purple arrow.) G) Local microenvironment may be estimated from spatial transcriptomic (ST) datasets by limiting cell-cell interactions to those within local neighborhoods, yielding a 'niche' atlas for each transcriptomic spot, which may be visualized in low dimensional space. H) Signaling mechanisms specific to the microenvironments of selected celltypes. Fgf1-Fgfr2 (cyan arrow) is a known potent regulator of oligodendrocyte phenotype (Furusho, et al., 2020; Furusho, et al., 2015) and here is found to be specific to the microenvironment of oligodendrocytes. I) Niche interactions of interest may be directly visualized in situ. J) Time-course data of cortical development in mouse captures neuron differentiation, including two distinct lineage trajectories. K) Pseudotemporal ordering of the same data shows the direction of differentiation. L) NICHES allows identification of key extracellular signals associated with lineage divergence over pseudotime, including known WNT- and NRXN-family signaling associated with cortical development (Jenkins, et al., 2016; Wang, et al., 2018)

Supplemental Figures

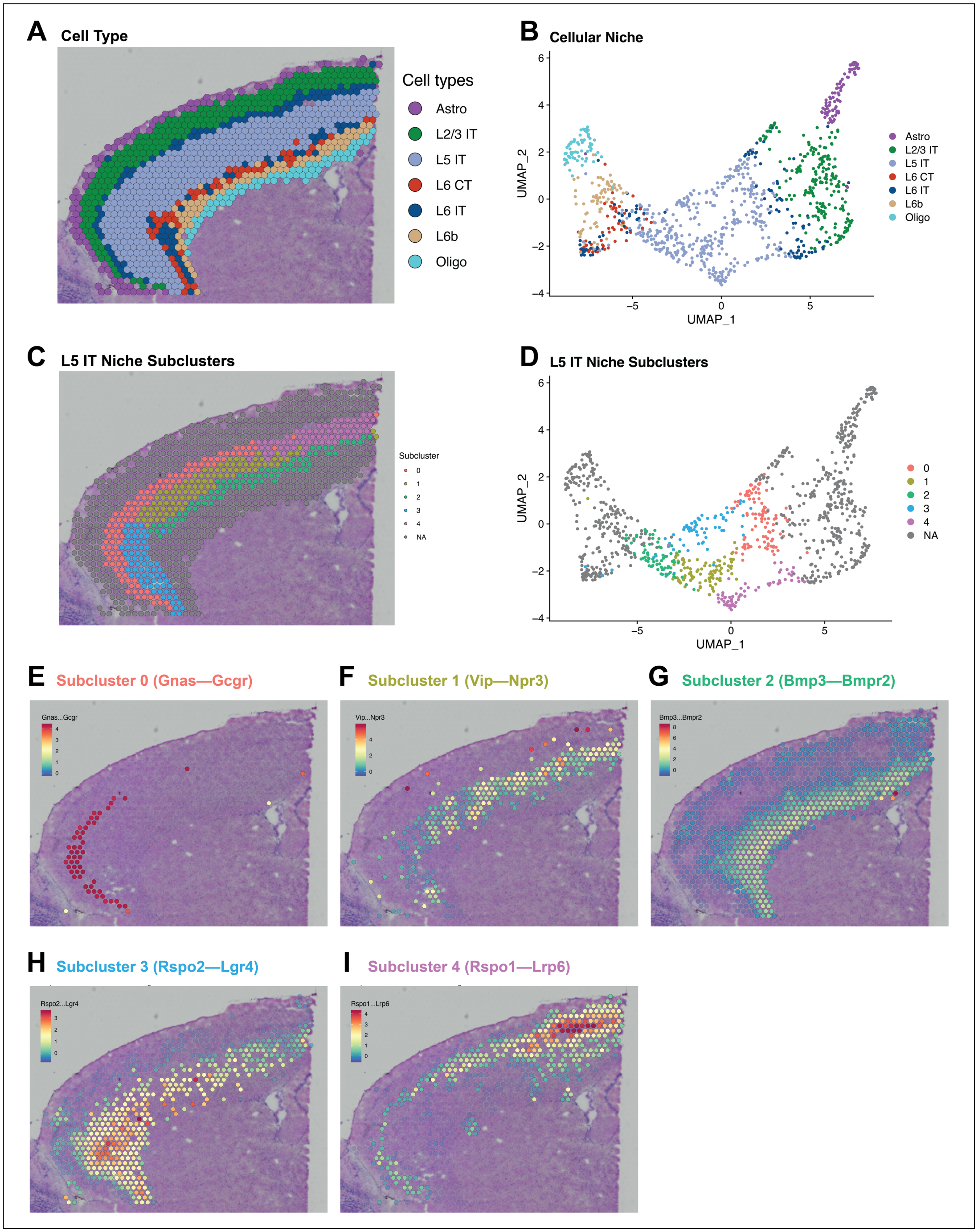


Supplemental Figure 1: NICHES captures biologically relevant nuances in cell-cell connectivity.

Simulation 1: NICHES reveals hidden intra-cluster heterogeneity (A-C.) In this first simulation, there are two celltypes labeled C1 and C2 (A, top). These two celltypes have hidden subpopulations communicating in distinct ways via two signaling mechanisms A2m-Lrp1 and Ace-Agtr2 (A, bottom). Subpopulation S2 expresses ligand A2m higher than S1 (B, top), while subpopulation S3 expresses receptor Agtr2 higher than S4 (B, bottom). This creates four distinct cell-cell signaling relationships even though only two celltypes have been crossed. When run through NICHES, all four distinct cellular relationships are clearly discernable in a two-dimensional embedding and may be grouped using k-means clustering (C, top). The k-means clusters almost perfectly resemble the ground truth subtype crosses (C, bottom). A black asterisk has been added showing the single connectivity value that would be calculated if only mean expression were used to compute connectivity between celltypes C1 and C2.

Simulation 2: NICHES preserves information regarding differential signaling distributions in disparate experimental conditions (E-I.) This simulation compares two cases (labeled Case 1 and Case 2), representing different experimental conditions with the same number of cells, same celltypes present, and similar mean connectivity (E). Case 1 sending cells express ligand sparsely but highly, while Case 2 sending cells express ligand broadly but lowly (F.) Receptor expression is identical in each Case (G). While mean connectivity is nearly identical (H), NICHES captures the significantly different connectivity between the two cases (I).

Simulation 3: NICHES allows high-dimensional visualization of altered Niche topology due to addition or removal of cells (J-L). This simulation demonstrates the capability of NICHES to represent altered system-cell signaling topology due to the addition or removal of cells. In Case 1, there are two communicating cell types C1 and C2 (J). In Case 2, a third cell type (C3) has entered the system which expresses ligands sensed by the other cells (K). The addition of signal from C3 alters the relevant ligand-receptor rows in the niche matrix associated with receiving population C2. When we use NICHES to calculate system-to-cell signaling within each case, we see a clear shift in the character of the sensed environment of celltype C2. NICHES therefore empowers the study of complex biological questions, such as how added, aberrant or infiltrating cells might affect the microenvironment of a receiving celltype across experimental conditions or disease states.



Supplemental Figure 2: NICHES Reveals Intra-Celltype Microenvironment Heterogeneity
A) Spatial transcriptomic data labeled by dominant celltype (see Methods). B) UMAP embedding of cellular niche for each transcriptomic location. C) Sub-clustering of the L5 IT niche

represented spatially and D) within UMAP space. Exploration of marker mechanisms reveals niche interactions specific to the microenvironments within each subcluster (E-I).

Supplemental Text

NICHES reveals hidden intra-cluster signaling heterogeneity

To demonstrate NICHES's ability to discover intra-cluster heterogeneity, we generated a synthetic single-cell RNA-seq dataset with 400 cells equally divided into 2 cell types (C1 and C2), which can be separated by their marker genes (Supp Fig A, top).

Additionally, there are 2 subtypes S1 and S2 within C1 and 2 subtypes S3 and S4 within C2 (Supp Fig A, bottom). By design, S1 and S2 interact differentially with S3 and S4 through 2 ligand-receptor mechanisms. Specifically, S2 has a higher expression of *A2m* ligand gene compared to S1 (corresponding receptor gene is *Lrp1* that is highly expressed in S3 and S4), and S3 has a higher expression of receptor gene *Agtr2* compared to S4 (corresponding ligand gene is *Ace* which is highly expressed in S1 and S2), as shown in Supp Fig B. Due to the subtle gene expression difference, one cannot distinguish the subtypes based on the overall gene expression profiles. But because the nuance involves ligand and receptor genes, we may apply NICHES to capture this signal.

Supp Fig C shows the cell-cell matrix output of NICHES, embedded in 2D mechanism space, clustered via k-means (top) and labeled by ground-truth subtype interaction (bottom). Sending cells are all from C1 and receiving cells are all from C2, and the space is 2-dimensional because there are only 2 ligand-receptor mechanisms in this simulation: *A2m-Lrp1* and *Agtr2-Ace*. Note that the clustering results (Supp Fig C, top) are almost completely identical to the ground truth subtype interaction pair labels (Supp Fig C, bottom). For reference, we have added a single black asterisk representing the single mean connectivity value between C1 and C2 which would result from using a mean-wise computational method to assess ligand-receptor connectivity.

NICHES preserves information regarding differential signaling distributions in disparate experimental conditions

Next, we sought to demonstrate that NICHES can register cellular archetype shift. We simulated 2 scRNAseq cases, both of which contain 2 cell types (C1 and C2), as shown in Supp Fig E. In both cases, we simulated 1 active ligand-receptor channel between C1 and C2 (C1 expresses the ligand gene A2m and C2 expresses the receptor gene Lrp1). By design, the C2 expression level of receptor Lrp1 is identical in both cases (Supp Fig G). C1 expression level of ligand A2m is sparse but high in Case 1, and broad but low in Case 2, with similar mean values. Because the mean expression level of A2m is similar, the mean connectivity of A2m-Lrp1 is similar between the 2 cases (Supp Fig H). NICHES, however, is easily able to detect this significant difference in connectivity (Supp Fig I).

NICHES allows high-dimensional visualization of altered systems-cell topology due to addition or removal of cells

Next, we sought to evaluate NICHES's ability to map changes in system-to-cell signaling topology. We simulated two datasets: the first (Case 1) has 2 cell types (C1 and C2, Supp Fig J) while the second (Case 2) has 3 cell types (C1, C2, and C3, Supp Fig K). C1 and C2 are connected via two mechanisms, A2m-Lrp1 and Ace-Agtr2, where C1 cells express the ligands A2m and Ace and C2 cells express the receptors Lrp1 and Agtr2. C3 also expresses ligands A2m and Ace. From the perspective of C2, the expression of ligand within this cell system is markedly increased when C3 is added (from a biological perspective, C2 has a higher 'chance' of receiving signal when the C3 population is present), a phenomenon which may be directly captured and visualized using NICHES (Supp Fig L).

Supplemental Methods

Mathematical Formalism

First, we define basic notations:

Table S1. Notations

Notation	Description and terminology
$\boldsymbol{C}$	$\boldsymbol{C} = \{c_1, c_2, \dots, c_{N_C}\}$. An ordered set of cells in the system where N_C is the total number of cells in the system. In the spatial transcriptomic datasets, we also use $\boldsymbol{C}$ to represent each measurement (e.g. spots).
$\boldsymbol{X}$	Normalized Gene Expression Matrix. For Cell c_i, $\boldsymbol{X}_{c_i} = \left[x_{c_i}^{g_1}, x_{c_i}^{g_2}, \dots, x_{c_i}^{g_{N_G}}\right]^T$ is its gene expression vector, where g_i is the ith gene $i \in \{1,2, \dots, N_G\}$ and where N_G is the total number of genes. For instance, $x_{c_i}^{l_k}$ is the gene expression level of ligand l_k from Cell c_i, $x_{c_j}^{r_t}$ is the gene expression level of receptor r_t from Cell c_j
$\boldsymbol{M}$	$\boldsymbol{M} = \{m_1, m_2, \dots, m_{N_M}\}$. An ordered set of ligand-receptor mechanisms where N_M is the total number of mechanisms. Each mechanism $m_k$has a corresponding ligand l_k and receptor r_k
$\boldsymbol{L}$	$\boldsymbol{L} = [l_1, l_2, \dots, l_{N_M}]$. A vector of reference ligands. Each l_k has a corresponding mechanism m_k from $\boldsymbol{M}$ in which it participates. l_k can consist of multiple subunits.
$\boldsymbol{R}$	$\boldsymbol{R} = [r_1, r_2, \dots, r_{N_M}]$. A vector of reference receptors. Each r_k has a corresponding mechanism m_k from $\boldsymbol{M}$ in which it participates. r_k can consist of multiple subunits.
$\boldsymbol{E}$	$\boldsymbol{E} \in \{0,1\}$ **,** unweighted and directed adjacency matrix that indicates which cells are connected and can interact in the system. For instance, $E_{ij} = 1$ indicates that Cell i and Cell j are connected and the ligand signals of Cell i can be received by the receptors of Cell j

Given the gene expression data $\boldsymbol{X}$ of a cell system $\boldsymbol{C}$, along with a list of known ligand-receptor mechanism $\boldsymbol{M}$, we aim to define a vector $S_{C_iC_j} \in \mathbb{R}^{N_M}$ for every connected cell pair cell i and cell j in a pre-defined cell adjacency matrix $\boldsymbol{E}$, such that $S_{C_iC_j}$ can characterize the N_M -dimensional ligand-receptor interaction profiles between cell i sending signal via ligand and cell j receiving signal via its receptors.

Cell-Cell Matrix

To construct the Cell-Cell Matrix, we define $S_{C_iC_j} = \left[s_{C_iC_j}^{m_1}, s_{C_iC_j}^{m_2}, \dots, s_{C_iC_j}^{m_{N_M}}\right]^T$ in which $s_{C_iC_j}^{m_k} = x_{C_i}^{l_k} \times x_{C_j}^{r_k}$, where the mechanism m_k consists of ligand l_k and receptor r_k. We choose the multiplication operation so when the ligand or the receptor are not expressed the product is zero, representing no cell-cell signaling.

We concatenate the $S_{C_iC_j}$ Cell-Cell Interaction vectors to construct the Cell-Cell Matrix: $\boldsymbol{S} \in \mathbb{R}^{N_M \times N_E}$, where N_M is the total number of mechanisms and $N_E = \sum_{i,j}^{N_C} E_{ij}$ is the total number of (directed) connected cells.

$\boldsymbol{S}$ can be used as input to many computational analysis pipelines, including dimensionality reduction, clustering, differential expression, pseudo-temporal ordering, and trajectory inference, etc., to investigate cell-cell interactions at the individual cell level.

Computing $\mathbf{E}$ for single-cell RNA-seq/spatial transcriptomic datasets

One step before computing a Cell-Cell Matrix is to compute the adjacency matrix $\boldsymbol{E}$. For single-cell RNA-seq datasets we assume a fully connected cellular system. However, the computational complexity of $\boldsymbol{S}$ becomes $\boldsymbol{O\left(N_C^2\right)}$**,** which greatly hinders the application of Cell-Cell Matrix onto cellular systems of large number of cells (e.g. $N_C > 1 \times 10^3$).

To reduce the complexity, we adopt a random sampling scheme to down-sample edges and to compute a new $\widetilde{\boldsymbol{E}}$ as follows: Let's denote the set of cell type labels in the system by $P = \{p_1, p_2, \dots, p_{N_P}\}$ where N_P is the total number of cell types. The set of cells associated with each type is denoted by $N = \{n_1, n_2, \dots, n_{N_P}\}$. For each pair of cell types within $\{(p_k, p_m) | k = 1,2, \dots, N_P; m = 1,2, \dots, N_P\}$, we draw 2 sets of cells $C^{sub,p_k}$and C^{sub,p_m} from cells of cell type p_k and p_m uniformly, i.e., $\{C^{sub,p_k} \subseteq C^{p_k} | \, |C^{sub,p_k}| = \min(n_k, n_m)\}$ and $\{C^{sub,p_m} \subseteq C^{p_m} | \, |C^{sub,p_m}| = \min(n_k, n_m)\}$ ($|S|$ denotes the number of elements in set S). Then we pair up C^{sub,p_k} and C^{sub,p_m}: $Q = \{(c_i^{sub,p_k}, c_i^{sub,p_m}) | i = 1,2, \dots, \min(n_k, n_m)\}$. Lastly, each entry in the new adjacency matrix can be set as $\tilde{E}_{ij} = \begin{cases} 1, & if\ (c_i, c_j) \subseteq Q \\ 0, & else \end{cases}$.

For spatial transcriptomic datasets we constrain the interactions among cells to be only within a certain distance threshold or a certain local neighborhood. To be more specific, let us denote $\boldsymbol{D}$ as the Euclidean distance matrix among cells computed from the spatial locations, where d_{ij} is the distance between Cell i and Cell j. Given a distance threshold r, E_{ij} is computed as $E_{ij} = \begin{cases} 1, & if\ d_{ij} \leq r \\ 0, & else \end{cases}$ for each entry of $\boldsymbol{E}$ for spatial transcriptomic datasets. Alternatively, the user can specify the parameter k which computes a k-nearest neighbor (knn) graph from $\boldsymbol{D}$ and the adjacency matrix E will be computed as a mutual nearest neighbor graph from this knn graph, i.e., as $E_{ij} = \begin{cases} 1, & if\ i,j\ are\ mutual\ neighbors \\ 0, & else \end{cases}$

Niche Matrix

Besides the base cell-cell interaction formulation, we extend our original definition of the Cell-Cell Matrix to investigate cellular niche and cellular influence interactions.

Specifically, we define the Niche Matrix as: $\boldsymbol{Y} \in \mathbb{R}^{N_M \times N_C}$, where N_M is the total number of mechanisms and N_C is the total number of cells in the system. A column vector of $\boldsymbol{Y}$ is defined as $Y_{C_i} = \left[y_{C_i}^{m_1}, y_{C_i}^{m_2}, \dots, y_{C_i}^{m_{N_M}}\right]^T \in \mathbb{R}^{N_M}$, i.e., each column of $\boldsymbol{Y}$ is a N_M -dimensional vector that characterizes the interaction profiles between cells sending ligand signals to Cell i which possesses the relevant receptors to receive these signals.

The connectivity value on one mechanism (e.g. m_k) between sending cells and Cell i is defined as $y_{C_i}^{m_k} = op(X^{l_k}) \times x_{C_i}^{r_k}$ where the mechanism m_k consists of ligand l_k and receptor r_k, X^{l_k} denotes the row vector of l_k's expression levels across the cells that are connected to Cell i, and $op()$ is a vector operator which, in our implementation, can be either sum (default) or $mean$.

Similarly, we define the Influence Matrix as: $\boldsymbol{Z} \in \mathbb{R}^{N_M \times N_C}$ where each column of $\boldsymbol{Z}$ is a N_M -dimensional vector that characterizes the interaction profiles between Cell i that sends the ligand signals and the cells receiving from it. Each connectivity value between Cell i and the system is defined as $z_{C_i}^{m_k} = x_{C_i}^{l_k} \times op(X^{r_k})$ where the mechanism m_k consists of ligand l_k and receptor r_k, X^{r_k} denotes the row vector of r_k's expression levels across the cells that connect to Cell i in the system, and $op()$ is again either sum (default) or $mean$.

For single-cell RNA-seq datasets without spatial coordinates, we assume a fully connected $\boldsymbol{E}$ involving all cells measured within a system. For spatial transcriptomic datasets, we construct $\boldsymbol{E}$ in the same fashion as for the spatial Cell-Cell Matrix, limiting edges to neighbors within radius r or within a user-defined set of nearest neighbors.

Metadata transfer and flexible differential analysis

NICHES allows the researcher to carry over metadata (i.e. sample labels, coarse- and fine-grain cluster labeling, experimental conditions) from source data, allowing rapid downstream differential analysis between already tagged groupings of cells. For every input metadata category, the NICHES Cell-Cell Matrix output object has Sending Metadata, Receiving Metadata, and Sending-Receiving Metadata associated with every column. The Niche Matrix and Influence Matrix have only Receiving Metadata and Sending Metadata associated with their columns, respectively.

Because each Cell-Cell Matrix contains many individual measurements of cell pairings (or environment-cell pairings in the Niche Matrix), differential analysis can be used to reveal ligand-receptor mechanisms preferential to a given celltype-celltype cross within a system, or to identify top differential signaling mechanisms across subject, disease state, experimental condition, or tissue. Such calculations may be performed for specific celltype-celltype crosses or for other custom groupings as the user desires, based on mapped metadata. We recommend using ROC analysis to measure how well a mechanism differentiates two groups compared to standard two-sample tests when the columns in Cell-Cell Matrix or Niche Matrix are no longer independent.

Considerations regarding data imputation

With unimputed data, the partial transcriptome sampling ('dropout') which is intrinsic to some single-cell RNA-seq technologies can cause some heterogeneity in downstream cell-to-cell measurements and resulting clustering due to which specific barcodes are paired. This phenomenon can occur even within a single celltype-to-celltype vector, which some users may consider artifactual and others may consider biologically relevant. Imputation greatly lessens this heterogeneity.

Application to simulated data

We generate 3 simulation datasets (*Simulation 1*, *Simulation 2*, *Simulation 3*) for 3 separate simulation analyses (Supplemental Fig A-C, Fig E-I, Fig J-L) respectively. For each dataset, we simulate 3 categories of genes: signaling genes (ligands and receptors), 50 non-signaling marker genes to differentiate each cell type, and 5000 noise genes. We assume the genes in the datasets follow negative binomial (NB) distributions parametrized by parameter μ which characterizes the mean expression level, and the dispersion parameter γ. We describe the exact parameter settings for each dataset as follows:

Table S2: Count matrix design for *Simulation 1*

	Cell type 1 (C1) (200 cells)		Cell type 2 (C2) (200 cells)	
	Subtype 1 (S1) (100 cells)	Subtype 2 (S2) (100 cells)	Subtype 3 (S3) (100 cells)	Subtype 4 (S4) (100 cells)
A2m	NB(μ=1, γ=20)	NB(μ=30, γ=20)	0	
Lrp1	0		NB(μ=30, γ=20)	NB(μ=30, γ=20)
Ace	NB(μ=30, γ=20)	NB(μ=30, γ=20)	0	
Agtr2	0		NB(μ=30, γ=20)	NB(μ=1, γ=20)
Marker genes (50 genes)	NB(μ=10, γ=20)		NB(μ=20, γ=20)	
Noise genes (5000 genes)	NB(μ=15, γ=20)			

Table S3: Count matrix design for *Simulation* 2 (Case 1)

	Cell type 1 (C1) (100 cells)		Cell type 2 (C2) (100 cells)
A2m	NB(μ=100, γ=20) (10 cells)	0 (90 cells)	0
Lrp1	0		NB(μ=5, γ=20)
Marker genes (50 genes)	NB(μ=10, γ=20)		NB(μ=20, γ=20)

Noise genes (5000 genes)	NB(μ=15, γ=20)

Table S4: Count matrix design for *Simulation* 2 (Case 2)

	Cell type 1 (C1) (100 cells)	Cell type 2 (C2) (100 cells)
A2m	NB(μ=10, γ=20)	0
Lrp1	0	NB(μ=5, γ=20)
Marker genes (50 genes)	NB(μ=10, γ=20)	NB(μ=20, γ=20)
Noise genes (5000 genes)	NB(μ=15, γ=20)	

Table S5: Count matrix design for *Simulation* 3

	Cell type 1 (C1) (1000 cells)	Cell type 2 (C2) (1000 cells)	Cell type 3 (C3) (1000 cells)
A2m	NB(μ=5, γ=20)	0	NB(μ=5, γ=20)
Lrp1	0	NB(μ=30, γ=20)	0
Ace	NB(μ=5, γ=20)	0	NB(μ=5, γ=20)
Agtr2	0	NB(μ=30, γ=20)	0
Marker genes (50 genes)	NB(μ=10, γ=20)	NB(μ=20, γ=20)	NB(μ=30, γ=20)
Noise genes (5000 genes)	NB(μ=15, γ=20)		

Application to native lung single-cell transcriptomic data

Data was downloaded from (Raredon, et al., 2019), subset to 4 main populations of interest, and run through standard principle component analysis (PCA), clustering, and UMAP embedding pipelines in Seurat (McInnes, et al., 2018; Stuart, et al., 2019). Data was imputed using ALRA (Linderman, et al., 2022) and then run through the NICHES function RunCellToCell. The resulting signaling matrix was then used to create a new Seurat object which was scaled and run through PCA again and embedding using UMAP. FindAllMarkers was used in Seurat to identify cell-cell interaction markers of interest.

Application to brain spatial transcriptomic data

Anterior mouse brain data was downloaded from 10x Genomics (2020) and preprocessed following the steps in Seurat (Stuart, et al., 2019), with subsetting to the frontal cortex region only. We integrated the data with a reference single-cell RNA-seq dataset (Tasic, et al., 2016) and used its cell type annotations to predict the labels of the spatial pixels by a probabilistic classifier (Seurat TransferData function). We then annotated each spatial pixel by its most probable label.

For NICHES matrix construction, we imputed the data with ALRA (Linderman, et al., 2022) based on the normalized data matrix, and then applied the NICHES Neighborhood-to-Cell function to compute niche signaling between direct histologic neighbors. The resulting niche matrix was embedded using UMAP (McInnes, et al., 2018) in Seurat 4.0 (Hao, et al., 2021). FindAllMarkers in Seurat was used to compute top markers for each celltype niche.

Application to pseudotemporally ordered cortical single-cell data

Single-cell RNA-seq data was taken from Di Bella, *et al*.(Di Bella, et al., 2021). Data was subset to the CPN and CFuPN lineages which included the apical progenitor, intermediate progenitor, migrating neuron, CPN, SCPN, and CThPN cell types. Timepoints E10.5, E11.5, and E12.5 were removed since they did not adequately capture the cell types of interest for downstream trajectory analysis. Niche signaling was computed for each experimental batch using the NICHES System-to-Cell function.

For the pseudotime analysis of the single-cell data, Louvain clustering was first conducted. A PAGA graph was computed to identify the branch point of the trajectories.(Wolf, et al., 2019) Finally, diffusion pseudotime was used to compute exact pseudotime values along the trajectories.(Haghverdi, et al., 2016) A random cell from the E13 apical progenitors was used as the root cell.

The data was subset to the CFuPN lineage and the mechanisms in the niche signaling data were subset to those that had at least 20% of non-zero values. Pearson correlation was computed between niche signaling and pseudotime for every cell and mechanisms of interest were identified using an FDR cutoff of 0.05 and a minimum correlation of 0.4. Niche matrix values for these mechanisms were then plotted in a heatmap ordered by the pseudotime values.

References

Baccin, C., *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature cell biology* 2020;22(1):38-48.

Browaeys, R., Saelens, W. and Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods* 2019:1-4.

Butler, A., *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* 2018;36(5):411.

Cabello-Aguilar, S., *et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research* 2020;48(10):e55-e55.

Davidson, S., *et al.* Single-cell RNA sequencing reveals a dynamic stromal niche that supports tumor growth. *Cell reports* 2020;31(7):107628.

Di Bella, D.J., *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* 2021;595(7868):554-559.

Efremova, M., *et al.* CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols* 2020;15(4):1484-1506.

Furusho, M., *et al.* Developmental stage ‐ specific role of Frs adapters as mediators of FGF receptor signaling in the oligodendrocyte lineage cells. *Glia* 2020;68(3):617-630.

Furusho, M., *et al.* Fibroblast growth factor signaling in oligodendrocyte ‐ lineage cells facilitates recovery of chronically demyelinated lesions but is redundant in acute lesions. *Glia* 2015;63(10):1714-1728.

Haghverdi, L., *et al.* Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* 2016;13(10):845-848.

Hao, Y., *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;184(13):3573-3587.e3529.

Jenkins, A.K., *et al.* Neurexin 1 (NRXN1) splice isoform expression during human neocortical development and aging. *Molecular psychiatry* 2016;21(5):701-706.

Jin, S., *et al.* Inference and analysis of cell-cell communication using CellChat. *bioRxiv* 2020.

Linderman, G.C., *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications* 2022;13(1):192.

McCarthy, N., Kraiczy, J. and Shivdasani, R.A. Cellular and molecular architecture of the intestinal stem cell niche. *Nature Cell Biology* 2020;22(9):1033-1041.

McInnes, L., Healy, J. and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* 2018.

Nabhan, A.N., *et al.* Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* 2018;359(6380):1118-1123.

Noël, F., *et al.* ICELLNET: a transcriptome-based framework to dissect intercellular communication. *BioRxiv* 2020.

Qadir, M.M.F., *et al.* Single-cell resolution analysis of the human pancreatic ductal progenitor cell niche. *Proceedings of the National Academy of Sciences* 2020;117(20):10876-10887.

Ramilowski, J.A., *et al.* A draft network of ligand–receptor-mediated multicellular signalling in human. *Nature communications* 2015;6:7866.

Raredon, M.S.B., *et al.* Single-cell connectomic analysis of adult mammalian lungs. *Science Advances* 2019;5(12):eaaw3851.

Raredon, M.S.B., *et al.* Connectome: computation and visualization of cell-cell signaling topologies in single-cell systems data. *bioRxiv* 2021.

Rodda, L.B., *et al.* Single-cell RNA sequencing of lymph node stromal cells reveals niche-associated heterogeneity. *Immunity* 2018;48(5):1014-1028. e1016.

Stuart, T., *et al.* Comprehensive integration of single-cell data. *Cell* 2019;177(7):1888-1902. e1821.

Tasic, B., *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience* 2016;19(2):335.

Tikhonova, A.N., *et al.* Cell-by-cell deconstruction of stem cell niches. *Cell stem cell* 2020;27(1):19-34.

Türei, D., *et al.* Integrated intra ‐ and intercellular signaling knowledge for multicellular omics analysis. *Molecular systems biology* 2021;17(3):e9923.

Tyler, S.R., *et al.* PyMINEr finds gene and autocrine-paracrine networks from human islet scRNA-Seq. *Cell reports* 2019;26(7):1951-1964. e1958.

Wang, Y., *et al.* Interplay of the Norrin and Wnt7a/Wnt7b signaling systems in blood–brain barrier and blood–retina barrier development and maintenance. *Proceedings of the National Academy of Sciences* 2018;115(50):E11827-E11836.

Wang, Y., *et al.* iTALK: an R package to characterize and illustrate intercellular communication. *BioRxiv* 2019:507871.

Wolf, F.A., Angerer, P. and Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* 2018;19(1):1-5.

Wolf, F.A., *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* 2019;20(1):59.

Zhang, Y., *et al.* CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic Acids Research* 2021.

Zhang, Y., *et al.* Cellinker: a platform of ligand–receptor interactions for intercellular communication analysis. *Bioinformatics* 2021.

Zhou, X., *et al.* Circuit design features of a stable two-cell system. *Cell* 2018;172(4):744-757. e717.