COMPARATIVE ANALYSIS OF CELL-CELL COMMUNICATION AT SINGLE-CELL RESOLUTION

Authors:
Aaron J. Wilk[1,2,3]*, Alex K. Shalek[4,5,6,7,8], Susan Holmes[9‡], and Catherine A. Blish[1,2,3,10‡]

Affiliations
[1]Stanford Immunology Program, [2]Department of Medicine, [3]Medical Scientist Training Program, Stanford University School of Medicine, Stanford, CA 94305, USA; [4]Institute for Medical Engineering & Science, [5]Department of Chemistry, and [6]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [7]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA; [8]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [9]Department of Statistics, Stanford University, Stanford, CA 94305, USA; [10]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA.

[‡]These authors jointly supervised this work
*To whom correspondence should be addressed: awilk@stanford.edu

Keywords: cell-cell communication, intercellular signaling, single-cell transcriptomics

ABSTRACT

Inference of cell-cell communication (CCC) from single-cell RNA-sequencing data is a powerful technique to uncover putative axes of multicellular coordination, yet existing methods perform this analysis at the level of the cell type or cluster, discarding single-cell level information. Here we present Scriabin – a flexible and scalable framework for comparative analysis of CCC at single-cell resolution. We leverage multiple published datasets to show that Scriabin recovers expected CCC edges and use spatial transcriptomic data to validate that the recovered edges are biologically meaningful. We then apply Scriabin to uncover co-expressed programs of CCC from atlas-scale datasets, validating known communication pathways required for maintaining the intestinal stem cell niche as well as previously unappreciated modes of intercellular communication. Finally, we utilize single-cell communication networks calculated using Scriabin to follow communication pathways that operate between timepoints in longitudinal datasets, highlighting bystander cells as important initiators of inflammatory reactions in acute SARS-CoV-2 infection. Our approach represents a broadly applicable strategy to leverage single-cell resolution data maximally toward uncovering CCC circuitry and rich niche-phenotype relationships in health and disease.

INTRODUCTION

Complex multicellular organisms rely on coordination within and between their tissue niches to maintain homeostasis and appropriately respond to internal and external perturbations. This coordination is achieved through cell-cell communication (CCC), whereby cells send and receive biochemical and physical signals that influence cell phenotype and function[1,2]. A fundamental goal of systems biology is to understand the communication pathways that enable tissues to function in a coordinated and flexible manner to maintain health and fight disease[3,4].

The advent of single-cell RNA-sequencing (scRNA-seq) has made it possible to dissect complex multicellular niches by applying the comprehensive nature of genomics at the “atomic” resolution of the single cell. Concurrently, the assembly of protein-protein interaction databases[5] and the rise of methods for pooled genetic perturbation screening[6,7] have empowered the development of methods to infer putative axes of cell-to-cell communication from scRNA-seq datasets[8–13]. These techniques generally function by aggregating ligand and receptor expression values for groups of cells to infer which groups of cells are likely to interact with one another[14–17]. However, biologically, CCC does not operate at the level of the group; rather, such interactions take place between individual cells. There exists a need for methods of CCC inference that: 1. analyze interactions at the level of the single cell; 2. leverage the full information content contained within scRNA-seq data by looking at up- and down-stream cellular activity; 3. enable comparative analysis between conditions; and, 4. are robust to multiple experimental designs.

Here we introduce single-cell resolved interaction analysis through binning (Scriabin) – an adaptable and computationally-efficient method for CCC analysis. Scriabin dissects complex communicative pathways at single-cell resolution by combining curated ligand-receptor interaction databases[13,18,19], models of downstream intracellular signaling[20], anchor-based dataset integration[21], and gene network analysis[22] to recover biologically meaningful CCC edges at single-cell resolution.

RESULTS

A flexible framework for comparative CCC analysis at single-cell resolution

Our goal was to develop a scalable and statistically robust method for the comprehensive analysis of CCC from scRNA-seq data. Scriabin implements three separate workflows depending on dataset size and analytical goals (**Figure 1**): 1. the cell-cell interaction matrix workflow, optimal for smaller datasets, analyzes communication methods used for each cell-cell pair in the dataset; 2. the summarized interaction graph workflow, designed for large comparative analyses, identifies cell-cell pairs with different total communicative potential between samples; and, 3) the interaction program discovery workflow, suitable for any dataset size, finds modules of co-expressed ligand-receptor pairs.

The fundamental unit of CCC is a sender cell N_i expressing ligands that are received by their cognate receptors expressed by a receiver cell N_j. Scriabin encodes this information in a

cell-cell interaction matrix ***M*** by calculating the geometric mean of expression of each ligand-receptor pair by each pair of cells in a dataset (**Figure 1A**). As ligand-receptor interactions are directional, Scriabin considers each cell separately as a "sender" (ligand expression) and as a "receiver" (receptor expression), thereby preserving the directed nature of the CCC network. ***M*** can be treated analogously to a gene expression matrix and used for dimensionality reduction, clustering, and differential analyses.

Next, Scriabin identifies biologically meaningful edges, which we define as ligand-receptor pairs that are predicted to result in observed gene expression profiles in the receiving cell (**Figure 1**). This requires defining a gene signature for each cell that reflects its relative gene expression patterns and determining which ligands are most likely to drive that observed signature. First, variable genes are identified across an axis of interest in order to immediately focus the analysis on features that distinguish samples of relevance or salient dynamics. When analyzing a single dataset, this set of genes could be the most highly-variable genes (HVGs) in the dataset, which would likely reflect cell type- or state-specific modes of gene expression. Alternatively, when analyzing multiple datasets, the genes that are most variable between conditions (or time points) could be used. To define the relationship between the selected variable genes and each cell, the single cells and chosen variable genes are placed into a shared low dimensional space with multiple correspondence analysis (MCA) implemented by Cell-ID[23], a generalization of principal component analysis (PCA). A cell's gene signature is defined as the set of genes in closest proximity to the variable genes in the MCA embedding (see **Methods**). NicheNet[20] is then used to nominate the ligands that are most likely to result in each cell's observed gene signature. Ligand-receptor pairs that are recovered from this process are used to weight the cell-cell interaction matrix ***M*** proportionally to their predicted activity, highlighting the most biologically important interactions (**Figure 1**).

Because one dimension of ***M*** is N x N cells long, it is impractical to construct ***M*** for samples with high cell numbers; this problem will likely be exacerbated as scRNA-seq platforms continue to increase in throughput. Conceptually, solutions to this problem include subsampling and aggregation. Subsampling, however, is statistically inadmissible because it involves omission of available valid data and introduction of sampling noise[24]; meanwhile, aggregation at any level raises the possibility of obscuring important heterogeneity and/or specificity.

A third solution is to first intelligently identify cell-cell pairs of interest and build ***M*** using only those sender and receiver cells. We hypothesize that, in the context of a comparative analysis, sender-receiver cell pairs that change substantially in their *magnitude* of interaction are the most biologically informative. To identify these cells, Scriabin first constructs a summarized interaction graph ***S***, characterized by an N by N matrix containing the sum of all cognate ligand-receptor pair expression scores for each pair of cells. ***S*** is much more computationally efficient to generate, store, and analyze than a full-dataset ***M*** (for a 1,000 cell dataset, ***S*** is 1,000 by 1,000, whereas ***M*** is ~3,000 by 1,000,000). Comparing summarized interaction graphs from multiple samples requires that cells from different samples share a set of labels or annotations denoting what cells represent the same identity. We use recent progress in dataset integration

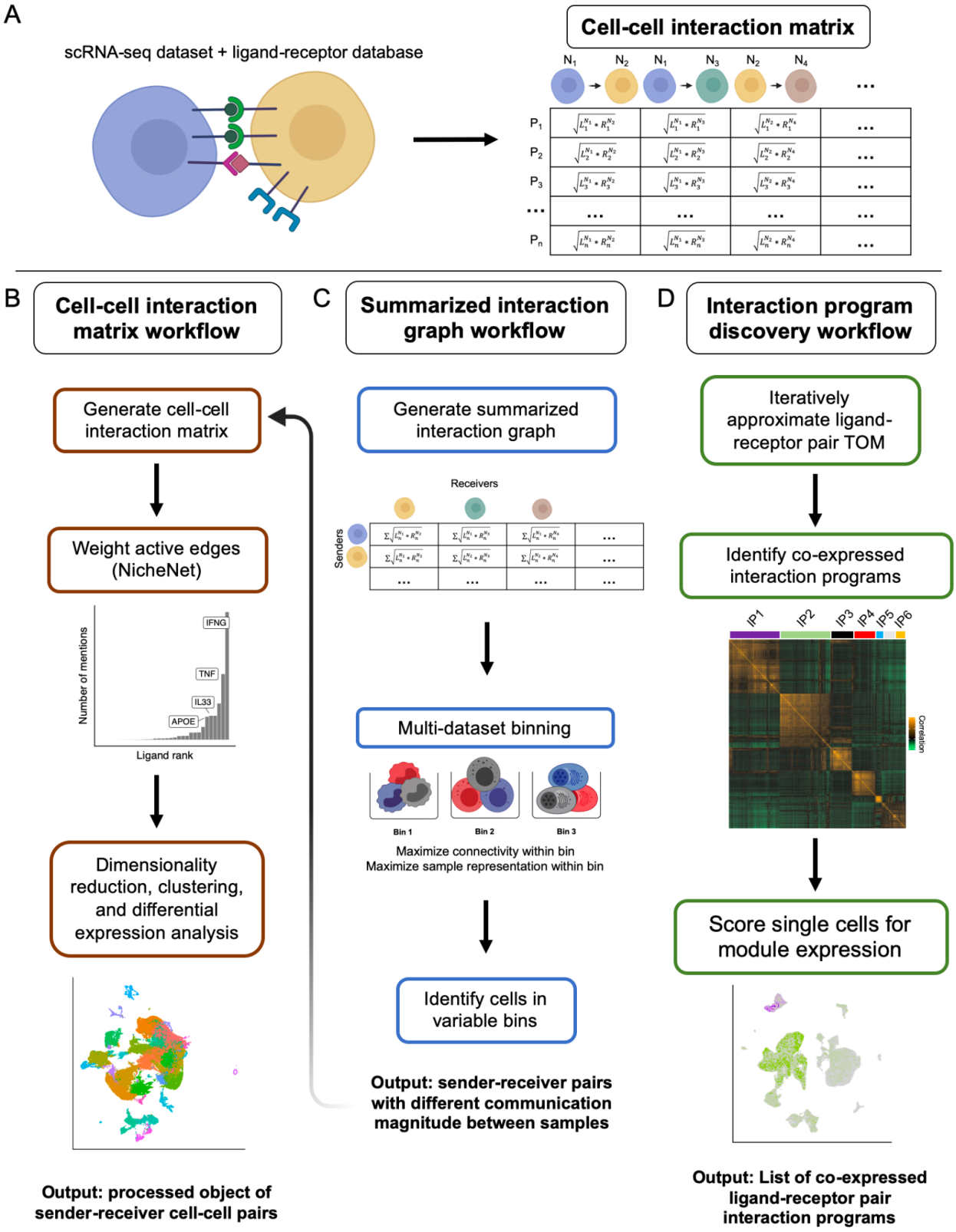


Figure 1: Schematic overview of cell-resolved communication analysis with Scriabin. Scriabin consists of multiple analysis workflows depending on dataset size and the user's analysis goals. **A)** At the center of these workflows is the calculation of the cell-cell interaction matrix ***M***, which represents all ligand-receptor expression scores for each pair of cells. **B)** Cell-cell interaction matrix workflow: In small datasets, ***M*** can be calculated directly, active CCC edges predicted using NicheNet[20], and the weighted cell-cell interaction matrix used for downstream analysis tasks like dimensionality reduction. ***M*** is a matrix of NxN cells by P ligand-receptor pairs, where each unique cognate ligand-receptor combination constitutes a unique P. **C)** Summarized interaction graph workflow: In large comparative analyses, a summarized interaction graph ***S*** can be calculated in lieu of a full-dataset ***M***. After high-resolution dataset alignment through binning, the most highly variable bins in total communicative potential can be used to construct an intelligently subsetted ***M***. **D)** Interaction program discovery workflow: Interaction programs of co-expressed ligand-receptor pairs can be discovered through iterative approximation of the ligand-receptor pair topological overlap matrix (TOM). Single cells can be scored for the expression of each interaction program (IP), followed by differential expression and modularity analyses.

methodology[21,25] to develop a high-resolution alignment process we call "binning," where we assign each cell a bin identity that maximizes the similarity of cells within each bin, maximizes the representation of all samples we wish to compare within each bin, while simultaneously minimizing the degree of agglomeration required (**Figure 1**; **Supplemental Text**). Sender and receiver cells belonging to the bins with the highest communicative variance can then be used to construct ***M***.

Finally, Scriabin implements a workflow for single cell-resolved CCC analysis that is scalable to any dataset size, enabling discovery of co-expressed ligand-receptor interaction programs. This workflow is motivated by the observation that transcriptionally similar sender-receiver cell pairs will tend to communicate through similar sets of ligand-receptor pairs. To achieve this, we adapted the well-established weighted gene correlation network analysis (WGCNA) pipeline[22] – designed to find modules of co-expressed genes – to uncover modules of ligand-receptor pairs that are co-expressed by the same sets of sender-receiver cell pairs, which we call "interaction programs". Scriabin calculates sequences of ***M*** subsets that are used to iteratively approximate a topological overlap matrix (TOM) which is then used to discover highly connected interaction programs. Because the dimensionality of the approximated TOM is consistent between datasets, this approach is highly scalable. The connectivity of individual interaction programs is then tested for statistical significance, which can reveal differences in co-expression patterns between samples. Single cells are then scored for the expression of statistically significant interaction programs. Comparative analyses include differential expression analyses on identified interaction programs, as well as comparisons of intramodular connectivity between samples.

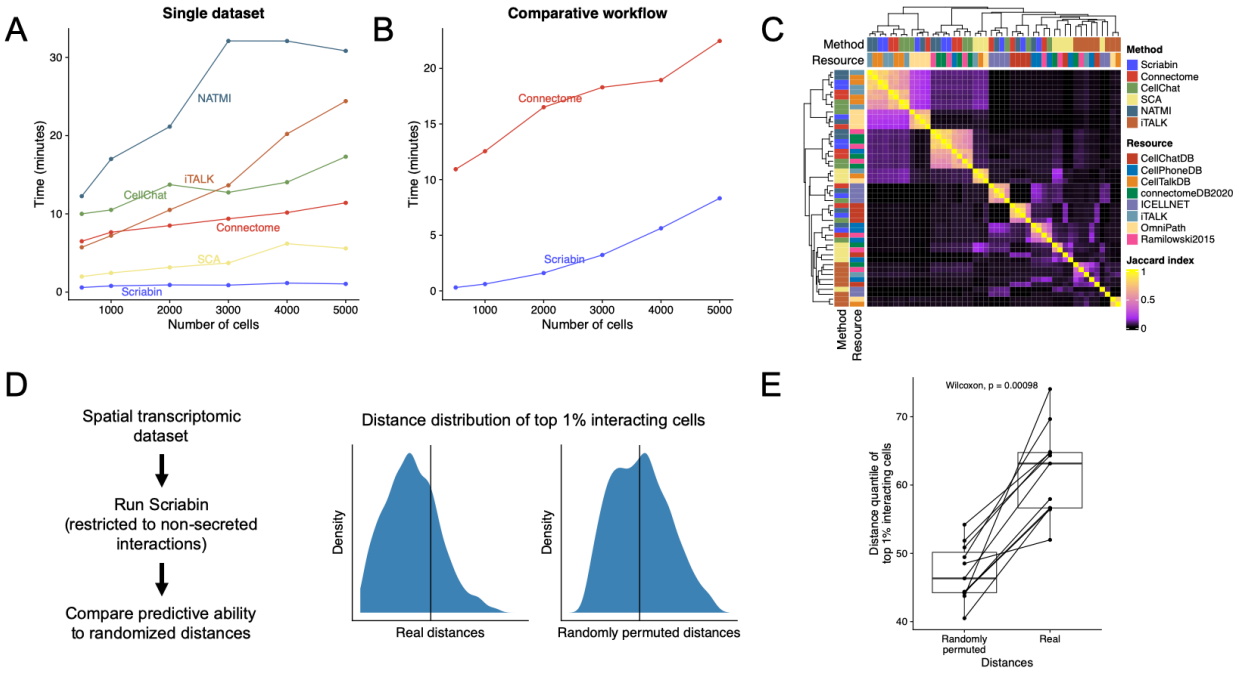


Figure 2: Benchmarking and robustness analysis of cell-resolved communication analysis. A) Runtime of Scriabin and five published CCC methods on the 10X PBMC 5k dataset. For each dataset size, the dataset was randomly subsampled to the indicated size and the same subsampled dataset was used for all methods. **B)** Runtime of Scriabin and Connectome comparative workflows. The 10X PBMC 5k and 10k datasets were merged into a single dataset which was subsampled as in (**A**), and the comparative workflows performed between cells from the 5k vs. 10k dataset. **C)** Jaccard index heatmap depicting the degree of overlap in the top 1,000 ligand-receptor CCC edges from each method-resource pair. **D)** Left, description of workflow to validate Scriabin using spatial transcriptomics datasets; right, density plots showing the distribution of cell-cell distances within the top 1% of highly interacting cell-cell pairs predicted by Scriabin. The vertical black lines denote the median distance of all cell-cell pairs. **E)** The procedure depicted in (**D**) was repeated for 11 datasets, and the median distance quantile of the top 1% interacting cell-cell pairs calculated using real cell distances relative to randomly permuted cell distances.

Scriabin is a robust and efficient method for single-cell resolved communication analysis

We next explored Scriabin's performance in comparison to other published CCC methods. Scriabin was faster than five agglomerative CCC methods[15–17,26,27] in analyzing a single dataset at all the dataset sizes tested (**Figure 2A**). Of these five agglomerative CCC methods, only Connectome[27] supports a full comparative workflow, and was slower than Scriabin in a comparative CCC analysis of two datasets (**Figure 2B**). We also compared the top CCC edges predicted by these methods[28] to a pseudobulk version of Scriabin, finding that the top results returned by Scriabin overlapped highly with three of the five published methods analyzed

(Connectome, CellChat, and NATMI; **Figure 2C**). The remaining two methods (iTALK and SCA) did not overlap with each other or any of the other tested methods (**Figure 2C**).

While Scriabin's results agreed with several published methods, we also sought to demonstrate more directly that these results were biologically correct. We hypothesized that spatial transcriptomic datasets could be leveraged for this purpose, as cells that Scriabin predicts to be highly interacting should be, on average, in closer proximity. We ran Scriabin on 11 spatial transcriptomic datasets, removing secreted ligand-receptor interactions that could operate over a distance from the ligand-receptor database (**Figure 2D**). Cells that Scriabin predicted were the most highly interacting were in significantly closer proximity relative to randomly permuted distances (**Figure 2E**).

Scriabin reveals known communicative biology concealed by agglomerative methods

We next evaluated if Scriabin's single-cell resolution CCC results returned communicating edges that are obscured by agglomerative CCC methods. To this end, we analyzed a publicly-available dataset of a well-characterized tissue niche: the granulomatous response to *Mycobacterium leprae* infection (**FIgure 3A**). Granulomas are histologically characterized by infected macrophages and other myeloid cells surrounded by a ring of Th1 T cells[29–31]. These T cells produce IFN-γ that is sensed by myeloid cells; this communication edge between T cells and myeloid cells is widely regarded as the most important interaction in controlling mycobacterial spread[32–34]. Ma et al. performed scRNA-seq on skin granulomas from patients infected with *Mycobacterium leprae*, the causative agent of leprosy[29]. This dataset includes granulomas from five patients with disseminated lepromatous leprosy (LL) and 4 patients undergoing a reversal reaction (RR) to tuberculoid leprosy, which is characterized by more limited disease and a lower pathogen burden (**Figure 3A**). Analysis of CCC with Scriabin revealed *IFNG* as the most important ligand sensed by myeloid cells in all analyzed granulomas, matching biological expectations (**Figure 3B**).

To assess if Scriabin was capable of avoiding pitfalls associated with agglomerative methods in comparative CCC analyses, we analyzed differential CCC pathways between LL and RR granulomas using an agglomerative method (Connectome; which implements a full comparative workflow[27]) and Scriabin. As Connectome performs differential CCC analyses by aggregating data at the level of cell type or cluster, it requires that each cluster have representation from the conditions being compared. In the Ma et al. dataset, satisfying this condition meant decreasing clustering resolution from 1 to 0.05 so that all major cell types are present in all profiled granulomas (**Supplementary Figure 1**) and comparing all aggregated LL granulomas to all aggregated RR granulomas (**Figure 3C**). This requirement moves analysis further from single-cell resolution and may allow individual donors to exert disproportionate influence on downstream analysis. Comparative CCC analysis with Connectome revealed *IL1B* and *CCL21* as the two most upregulated T cell-expressed ligands received by myeloid cells in RR granulomas (**Figure 3D**). However, there was no clear pattern of *IL1B* upregulation among RR granulomas (**Figure 3E**); rather, the RR granuloma that contributed the most T cells expressed

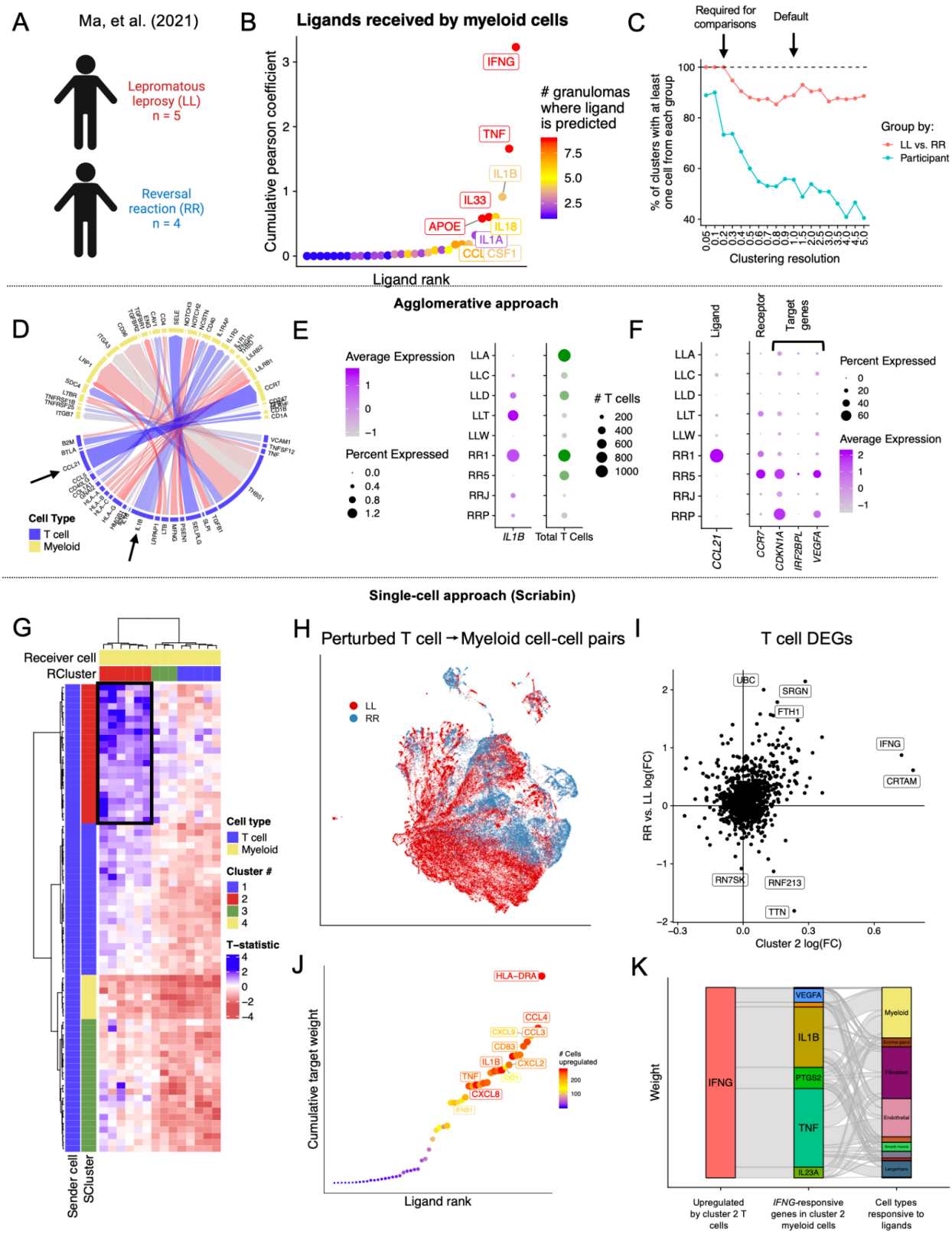


Figure 3: Scriabin reveals communicative pathways obscured by agglomerative techniques. A) Schematic of the scRNA-seq dataset of leprosy granulomas published by Ma, et al. **B)** Ligands prioritized by Scriabin's implementation of NicheNet as predicting target gene signatures in granuloma myeloid cells. Points are colored and sized by the number of granulomas in which the ligand is predicted to result in the downstream gene signature. **C)** Clustering resolutions required for comparative CCC analysis by agglomerative methods. Pink bars indicate the percentage of clusters containing at least one cell from an LL granuloma and one cell from an RR granuloma. Blue bars indicate the percentage of clusters containing at least one cell from all nine analyzed granulomas. **D)** Circos plot summarizing RR vs. LL differential CCC edges between T cells (senders) and myeloid cells (receivers) generated by Connectome. Blue, edges upregulated in RR; red, edges upregulated in LL. **E)** Percentage and average of expression of *IL1B* by T cells per granuloma (left), and total number of T cells per granuloma (right). **F)** Percentage and average expression of *CCL21* by T cells per granuloma (left); percentage and average expression of *CCR7* and *CCL21*-stimulated genes by myeloid cells per granuloma. **G)** RR vs. LL differential interaction heatmap between T cell bins (senders; rows) and myeloid cell bins (receivers; columns) generated by Scriabin. In blue, are bins more highly interacting in RR; in red are the bins more highly interacting in LL. The black box indicates groups of bins predicted to be highly interacting in RR granulomas relative to LL. **H)** UMAP projection of perturbed T cell-myeloid cell sender-receiver pairs indicating changes in ligand-receptor pairs used for T cell-myeloid communication in LL vs. RR granulomas. **I)** Scatter plot depicting differential gene expression by T cells. The average log(fold-change) of expression by cluster 2 bins is plotted on the x-axis; the average log(fold-change) of expression by RR granulomas is plotted on the x-axis. **J)** Target genes predicted to be upregulated by *IFNG* in RR granuloma myeloid cells in cluster 2 bins. Points are sized and colored by the number of cells in which the target gene is predicted to be *IFNG*-responsive. **K)** Alluvial plot depicting the RR granuloma cell types that are predicted to receive the *IFNG*-responsive target genes from cluster 2 myeloid cells.

the highest level of *IL1B* and the LL granuloma that contributed the fewest T cells expressed the lowest level of *IL1B* (**Figure 3E**). Additionally, *CCL21* was expressed by T cells of a single RR granuloma, and the myeloid cells of a different RR granuloma expressed the highest levels of the *CCL21* receptor *CCR7* and three *CCL21* target genes (**Figure 3F**). This indicates that the most highly scored differential CCC edges may likely be due to agglomeration of RR and LL granulomas required by Connectome (**Figure 3C**), rather than conserved biological changes between these two groups.

To compare differential CCC between LL and RR granulomas with Scriabin, we aligned data from the 9 granulomas together using Scriabin's binning procedure (**Figure 1**), generated single-cell summarized interaction graphs for each granuloma, and calculated a t-statistic to quantify the difference in interaction for each pair of bins between LL and RR granulomas

(**Figure 3G**). This analysis revealed a group of T cell and myeloid bins whose interaction was strongly increased in RR granulomas relative to LL (**Figure 3G**, black box). We visualized the cells in these perturbed bins by generating cell-cell interaction matrices for these cells in each sample and embedding them in shared low dimensional space (**Figure 3H**). The T cells in these bins were defined by expression of *CRTAM*, a marker of cytotoxic CD4 T cells, and upregulated *IFNG* in the RR granulomas (**Figure 3I**). Myeloid cells in these bins upregulated several pro-inflammatory cytokines in RR granulomas, including *IL1B, CCL3*, and *TNF* in response to *IFNG* from this T cell subset (**Figure 3J**). *IFNG*-responsive *IL1B* and *TNF* were also predicted to be RR-specific ligands received by myeloid cells, fibroblasts, and endothelial cells in RR granulomas (**Figure 3K**). Collectively, Scriabin identified a subset of *CRTAM*+ T cells that upregulated *IFNG* in RR granulomas that is predicted to act on myeloid cells to upregulate additional pro-inflammatory cytokines. These CCC results match previous results demonstrating that enhanced production of *IFNG* can drive RRs[35,36] and implicate cytotoxic CD4 T cells as initiators of this reaction.

Discovery of co-expressed interaction programs enables atlas-scale analysis of CCC at single-cell resolution

We next assessed Scriabin's interaction program discovery workflow. To illustrate the scalability of this process, we chose to analyze a large single-cell atlas of developing fetal gut[37] composed of ~80,000 cells sampled from four anatomical locations (**Figure 4A**). Scriabin discovered a total of 75 significantly correlated interaction programs across all anatomical locations. Scoring all single cells on the expression of the ligands and receptors in these interaction programs revealed strong cell type specific expression patterns for many programs (**Figure 4B**).

We next examined ways in which our identified interaction programs reflected known biology of intestinal development. Recently, several important interactions have been shown to be critical in maintaining the intestinal stem cell (ISC) niche[38–40]. We were able to identify ISCs, defined by expression of *LGR5* and *SOX9*, within the intestinal epithelial cells of this dataset, and discovered a single interaction program (hereafter referred to as IP1) whose receptors were co-expressed with these ISC markers (**Figure 4C**). IP1 represents a program of fibroblast-specific ligand and intestinal epithelial cell receptor expression (**Figure 4D**). Among IP1 ligands were the ephrins *EPHB3,* whose expression gradient is known to control ISC differentiation[41], and *RSPO3* (**Figure 4E**). Two recent studies have each reported that *RSPO3* production by lymphatic endothelial cells (LECs) and GREM1+ fibroblasts is critical for maintaining the ISC niche[39,40]. While we did not observe expression of *RSPO3* in LECs (**Supplementary Figure 2**), we did find that *GREM1*+ fibroblasts expressed *RSPO3* as a part of IP1 that was predicted to be sensed primarily by ISCs (**Figure 4D-F**). We also found a separate interaction program containing the ligand *GREM1*; the ligands of this interaction program were co-expressed with IP1 ligands (**Figure 4F**) and predicted to communicate to a different receiver cell type, namely gut endothelial cells (**Figure 4G**). Additionally, further investigation of cell type-specific modules revealed subtle within-cell type differences in sender or receiver potential,

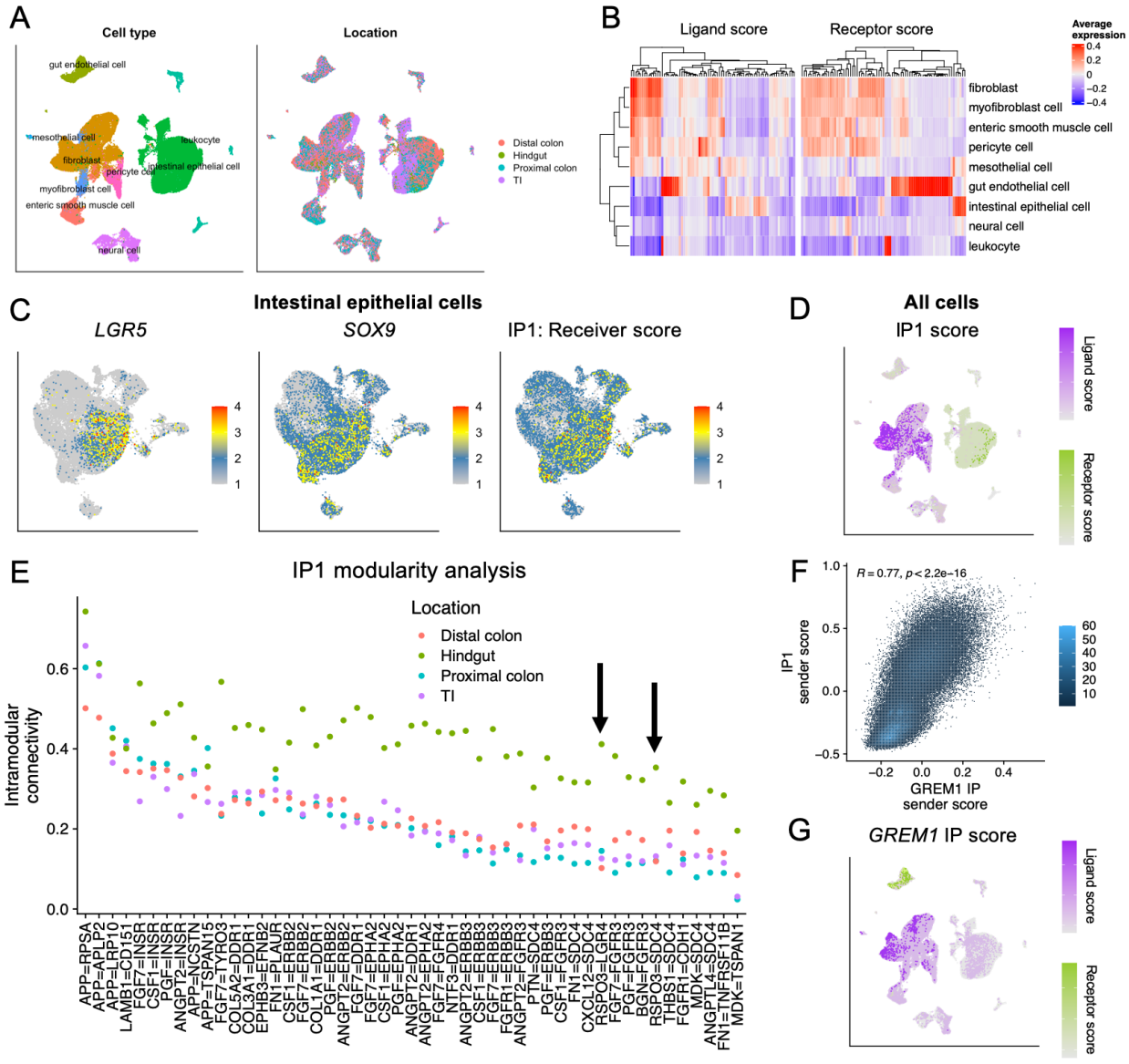


Figure 4: Cell-cell interaction programs of the developing fetal gut. A) UMAP projections of the dataset of Fawkner-Corbett, et al.[37], with individual cells colored by author-provided cell type annotations (left), or by anatomical sampling location (right). **B)** Heatmap depicting average expression of interaction program ligands (left) or interaction program receptors (right) by each cell type. **C)** UMAP projections of intestinal epithelial cells, colored by expression of stem cell markers *LGR5* and *SOX9*, as well as by the receptor expression score for interaction program 1 (IP1). **D)** UMAP projection of all cells colored by ligand (shades of purple) or receptor (shades of green) expression of IP1. **E)** Intramodular connectivity scores for each ligand-receptor pair in each anatomical location for IP1. The black arrows mark ligand-receptor pairs that include *RSPO3*. **F)** Heatmap of 2d bin counts depicting the correlation between IP1 sender score and the

sender score for the IP module that contains the ligand *GREM1*. **G)** UMAP projection of all cells colored by ligand (shades of purple) or receptor (shades of green) expression of the *GREM1* IP.

highlighting the importance of maintaining single-cell resolution (**Supplementary FIgure 2; Supplementary Text**). Taken together, our results suggest interaction program discovery enables scalable CCC analysis at single-cell resolution that is capable of identifying known CCC edges.

Assembly of longitudinal communicative circuits

A frequent analytical question in longitudinal analyses concerns how events at one time point influence cellular phenotype in the following time point. We hypothesized, in datasets with close spacing between time points, that Scriabin's high-resolution bin identities would allow us to assemble "longitudinal communicative circuits": chains of sender-receiver pairs across consecutive timepoints. A communicative circuit consists of at least four cells across at least two time points: sender cell at time point 1 (S_1), receiver cell at time point 1 (R_1), sender cell at time point 2 (S_2), and receiver cell at time point 2 (R_2). If the interaction between S_1-R_1 is predicted to result in the upregulation of ligand L_A by R_1, S_1-R_1-S_2-R_2 participates in a longitudinal circuit if R_1 and S_2 share the same bin (i.e., S_2 represents the counterpart of R_1 at timepoint 2) and if L_A is predicted to be an active ligand in the S_2-R_2 interaction (**Figure 5A**). This process enables the stitching together of multiple sequential timepoints to identify communicative edges that are downstream in time and mechanism.

To illustrate this process, we analyzed a published dataset of SARS-CoV-2 infection in human bronchial epithelial cells (HBECs) in air-liquid interface (ALI) that was sampled daily for 3 days[42]. This dataset contains all canonical epithelial cell types of the human airway and indicates that ciliated and club cells are the preferentially infected cell types in this model system, with some cells having >50% of UMIs from SARS-CoV-2 (**Figure 5B**). We first defined a per-cell gene signature of genes variable across time, and used this gene signature to predict active ligands expected to result in the observed cellular gene signatures[20,23]. Next, we used Scriabin's high-resolution binning workflow to align the datasets from the three post-infection timepoints, which we then used to assemble longitudinal communicative circuits.

Scriabin identified circuits at the level of individual cells that spanned all three post-infection timepoints. We summarized these circuits by author-annotated cell type and whether SARS-CoV-2 reads were detected in the cell (**Figure 5C**). We found that several acute-phase reactant-encoding genes, like *SAA1*, *CTGF*, and *C3* (encoding complement factor 3), were the most common ligands to participate in circuits across timepoints, matching biological expectations (**Figure 5C**)[43–45]. Interestingly, uninfected cells were more frequently the initiators of longitudinal circuits, producing some acute-phase reactant-encoding genes in addition to the pro-inflammatory cytokine gene *IL1B*. When we assessed the predicted downstream targets at the ends of the longitudinal circuits in both infected and bystander cells, we found that *TGFB1*

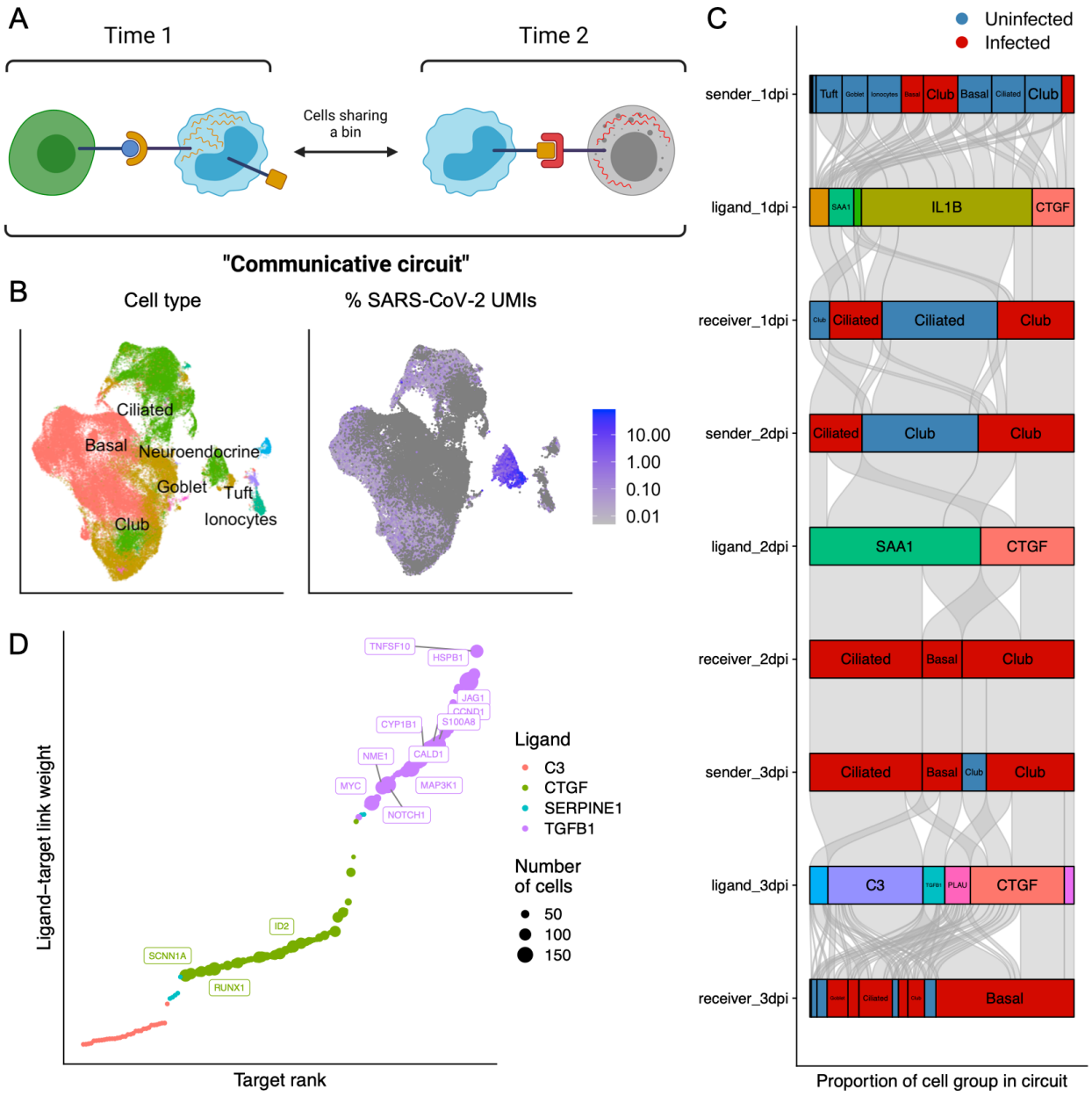


Figure 5: Longitudinal circuits of CCC in acute SARS-CoV-2 infection. A) Schematic representing a longitudinal communicative circuit. Four cells participate in a longitudinal circuit if an interaction between S_1-R_1 is predicted to result in the upregulation of ligand L_A by R_1, if R_1 and S_2 share a bin, and if expression of L_A by S_2 participates in an active communication edge with R_2. **B)** UMAP projections of the dataset published by Ravindra, et al.[42] colored by annotated cell type (left) or by the percentage of UMIs per cell of SARS-CoV-2 origin (right). **C)** Alluvial plot depicting longitudinal communicative circuits. Stratum width corresponds to the number of cells in each cell grouping participating in the circuit corrected for the total number of cells in that group. Red strata are infected

with SARS-CoV-2; blue strata are composed of uninfected cells. **D)** Target genes predicted by Scriabin's implementation of NicheNet[20] to be upregulated in the receiver cells at the ends of the longitudinal communicative circuits at 3 days post-infection. Points are colored by the active ligand and sized by the number of cells in which the target is predicted to be upregulated by the active ligand.

produced by infected basal cells was predicted to result in the upregulation of *TNFSF10* (encoding TRAIL) and the alarmin *S100A8* predominantly by other infected cells (**Figure 5C-D**). Additionally, *TGFB1* was predicted to upregulate both *NOTCH1* and the *NOTCH1* ligand *JAG1*, which indicates that these circuits may induce downstream Notch signaling. In sum, these data illustrate how the single-cell resolution of Scriabin's CCC analysis workflow can perform comprehensive and integrated longitudinal analyses.

DISCUSSION

Most existing CCC methodologies aggregate ligand and receptor expression values at the level of the cell type or cluster, potentially obscuring biologically valuable information. Here we introduce a framework to perform comparative analyses of CCC at the level of the individual cell. Scriabin maximally leverages the single-cell resolution of the data to maintain the full structure of both communicative heterogeneity and specificity. We used this framework to find rare communication pathways in the developing intestine known to be required for stem cell maintenance, as well as to define the kinetics of early dynamic communication events in response to SARS-CoV-2 infection through assembly of longitudinal communicative circuits.

A major challenge of single-cell resolved CCC analysis is data inflation: because CCC analysis fundamentally involves performing pairwise calculations on cells or cell groups, it is frequently computationally prohibitive to analyze every sender-receiver cell pair. Importantly, Scriabin implements two complementary workflows to address this issue, both of which avoid the statistically-problematic practices of subsampling and aggregation while maintaining scalability. Subsampling and aggregation preclude a truly comprehensive view of CCC structure and risk obscuring important biology; either can be particularly problematic in situations where a small subset of cells disproportionately drives intercellular communication, with agglomeration potentially concealing the full activity of those cells and subsampling potentially removing those cells altogether. One biological situation in which the preservation of single-cell resolution data could be particularly important is in the setting of activation-induced T cell exhaustion[46]. Exhausted T cells may be difficult to distinguish from activated cells by clustering or sub-clustering, yet exert divergent effects on their communication targets than their activated counterparts. By avoiding aggregation and subsampling, Scriabin increases the likelihood of detecting these potentially meaningful differences in CCC pathways.

We observe that aggregation obscures a potentially biologically-meaningful subset of T cells that arises during reversal reactions in leprosy granulomas. Due to the degree of transcriptional perturbation in T cells during reversal reactions, subclustering is not always a tenable approach

to increasing the resolution of CCC analyses because it, in turn, can preclude analysis at a per-sample level. We also show that aggregating across samples, a common practice in existing CCC tools, can return putatively differential CCC edges that are driven disproportionately by individual samples.

As the throughput of scRNA-seq workflows continues to increase, it will be important that single-cell resolution CCC methods are scalable to any dataset size. The interaction program discovery workflow of Scriabin accomplishes this by focusing first on common patterns of ligand-receptor pair co-expression rather than individual cell-cell pairs. Individual cells can be scored for expression of these interaction programs *post hoc*, enabling downstream comparative analyses with a comprehensive view of CCC structure. We apply this workflow to an atlas-scale dataset of human fetal gut development, where we validate a mode of communication between a fibroblast subset and ISCs that has recently been shown to be important for maintaining the ISC niche[39,40]. Due to the relative scarcity of these cells, it is likely that agglomerative methods may fail to discover these important interactions for downstream mechanistic validation.

Longitudinal datasets pose an additional opportunity and challenge for comparative analyses because there is *a priori* knowledge about the sequential relationship between different samples. The single-cell nature of Scriabin's workflows permits us to analyze how pathways of CCC operate both within and between timepoints in longitudinal datasets. By identifying circuits of CCC that function over multiple timepoints in a longitudinal infection dataset, we can observe how uninfected bystander cells may initiate important inflammatory pathways first which are later amplified by infected cells. A fundamental assumption of the circuit assembly workflow is that ligands upregulated at one timepoint can be observed to exert their biological activity at the following timepoint. This assumption is highly dependent on *a priori* biological knowledge of the communication pathways of interest, as well as on the spacing between timepoints. Assembly of longitudinal communication circuits may represent a valuable strategy to elucidate complex dynamic and temporal signaling events, particularly when longitudinal sampling is performed at frequencies on the same scale as signaling and transcriptional response pathways.

The cell-cell interaction matrix ***M*** is more highly enriched for zero values than gene expression matrices. This is because genes encoding molecules involved in CCC tend to be more lowly expressed than other genes (as the most highly expressed genes tend to encode intracellular proteins involved in cell housekeeping), and because a zero value in *either* the ligand or the receptor of a cell-cell pair will result in a zero value in the interaction vector. Consequently, these zero values can make it difficult for Scriabin to determine if putatively downregulated or "missing" CCC edges are biological or due to dropout. Data imputation or smoothing algorithms may be applied to the gene expression matrix and the resulting data used to generate ***M***. The high-resolution binning workflow implemented by Scriabin could also be used for data smoothing and prevent excessive propagation of zero values. This process can make the presence and absence of CCC edges more interpretable, but by smoothing the data may also slightly obscure underlying heterogeneity or structure. A potential middle ground solution to this

issue could be to perform data imputation for all features except the ligand-receptor pairs with the strongest expression values.

One current limitation of Scriabin is that it does not take into account situations where multiple receptor subunits encoded by different genes are required in combination to respond to a ligand, or where receptor subunits are known to differentially contribute to collective ligand-receptor avidity. An additional limitation is that Scriabin assumes uniform validity of ligand-receptor interactions in curated protein-protein interaction databases. Scriabin also treats all ligand-receptor pairs as equally important. In situations where it is known *a priori* which ligand-receptor pairs have a higher level of literature support, this information could be used to prioritize downstream analysis of particular ligand-receptor pairs. Similarly, all downstream signaling analyses in Scriabin rely on NicheNet's ligand-target activity matrix, which may be biased by the cell types and stimulation conditions used to generate it. The NicheNet database also does not allow for analysis of inhibitory signaling, and thus Scriabin will only return CCC edges predicted to result in activated signaling. While Scriabin uses NicheNet to predict active CCC edges by examining downstream gene expression changes, an additional analysis goal includes identifying the upstream signaling machinery that results in the upregulation of a ligand or denotes successful signaling, as additional power could be gained by using sets of genes to infer upstream signaling rather than relying on ligand expression alone (which could be impacted by dropout or differences between mRNA and protein expression). More generally, Scriabin assumes that gene expression values for ligands and receptors correlate well with their protein expression. A future point of improvement would be to support analysis of multi-modal datasets where cell surface proteins that contribute to CCC are measured directly, or to enable analysis of protein measurements that are imputed from integration with multi-modal references[47]. Future iterations of Scriabin will seek to address these issues, as well as further improve computational efficiency.

Collectively, our work provides a toolkit for comprehensive comparative analysis of CCC in scRNA-seq data, which should empower discovery of information-rich communicative circuitry and niche-phenotype relationships.

ACKNOWLEDGEMENTS

We thank Drs. William J. Greenleaf and Sam W. Kazer for helpful conversations in the conceptualization of Scriabin's workflow. We thank Constantine Tzouanas and Dr. José Ordovas-Montanes for insights on intestinal cell-cell communication pathways. We also thank all current and former members of the Blish laboratory for helpful discussions of this work. A.J. Wilk is supported by the Stanford Medical Scientist Training Program (T32 GM007365-44) and the Stanford Bio-X Interdisciplinary Graduate Fellowship. This work was supported by NIH/NIDA DP1 DA04508902 to C.A.B., NIH/NCI 1U54CA217377, U01 28020510, and 1U2CCA23319501 to A.K.S., NIH/NIDA 1DP1DA053731 to A.K.S., Bill and Melinda Gates Foundation INV-027498 and OPP1202327 to A.K.S., the MIT Stem Cell Initiative through Foundation MIT to A.K.S. and a 2019 Sentinel Pilot Project from the Bill and Melinda Gates Foundation to C.A.B./A.K.S..

AUTHOR CONTRIBUTIONS

A.J.W., A.K.S., S.H., and C.A.B. conceived of the work. A.J.W. built Scriabin, performed computational analyses, and wrote the manuscript with input from all authors. S.H. and C.A.B. jointly supervised the work.

DECLARATION OF INTERESTS

A.K.S. reports compensation for consulting and/or SAB membership from Merck, Honeycomb Biotechnologies, Cellarity, Repertoire Immune Medicines, Ochre Bio, Third Rock Ventures, Hovione, Relation Therapeutics, FL82, Empress Therapeutics, and Dahlia Biosciences. C.A.B. reports compensation for consulting and/or SAB membership from Catamaran Bio, DeepCell Inc., Immunebridge, and Revelation Biosciences.

DATA AND CODE AVAILABILITY

All scRNA-seq data analyzed in this manuscript are publicly available. Scriabin is available for download and use as an R package at github.com/BlishLab/scriabin

METHODS

Cell-cell interaction matrix workflow

Generation of cell-cell interaction matrix

We define the cell-cell interaction vector between a pair of cells as the geometric mean of expression values of each cognate ligand-receptor pair. Formally, the interaction vector V between sender cell N_i and receiver cell N_j is given by

$$V_{N_i N_j} = \left[\sqrt{N_i^{l_1} * N_j^{r_1}}, \sqrt{N_i^{l_2} * N_j^{r_2}}, \ldots, \sqrt{N_i^{l_n} * N_j^{r_n}} \right],$$

where l_n, r_n represent a cognate ligand-receptor pair. We chose to multiply ligand and receptor expression values so that zero values of either ligand or receptor expression would result in a zero value for the corresponding index of the interaction vector. Additionally, we chose to take the square root of the product of ligand-receptor expression values so that highly expressed ligand-receptor pairs do not disproportionately drive downstream analysis. This definition is equivalent to the geometric mean. The cell-cell interaction matrix ***M*** is constructed by concatenating the cell-cell interaction vectors. Linear regression is used to correct ***M*** for variation due to sequencing depth, where the total sequencing depth of $N_i N_j$ is defined as the sum of unique molecular identifiers (UMIs) in N_i and N_i. ***M*** is used as input to low dimensional embeddings for visualization, and nearest neighbor graphs for graph-based clustering.

Weighting cell-cell interaction matrix by upstream regulome

Cell-cell interaction matrix ***M*** can be weighted by ligand-receptor edges that are predicted to be active based on observed downstream gene expression changes. First, we identify genes in the dataset that are variable across some axis of interest. For analyses of single datasets, variable genes can be defined as the set of genes with the highest residual variance in the dataset, for example, by calling FindVariableFeatures as implemented by Seurat. For comparative analyses, Scriabin provides several utility functions to aid in the identification of variable genes between samples or between time points, depending on the user's analytical questions.

Next, the package CelliD[23], which provides a convenient and scalable workflow to define single-cell gene signatures, is used to define per-cell gene signatures. Briefly, user-defined variable genes are used to embed the dataset into low dimensional space by multiple correspondence analysis (MCA). A cell's gene signature is then defined as the set of genes to which that cell is nearest in the MCA biplot. A quantile cutoff is used to threshold gene proximity, by default the 10% of nearest genes.

NicheNet[20] is then used to rank ligands based on their predicted ability to result in the per-cell gene signature. First, expressed genes are defined by the percentage of cells in which they are detected (by default, 2.5%). Next, a set of potential ligands is defined as those ligands which are expressed genes and for which at least one receptor is also an expressed gene. Next, ligand

activities are predicted by predict_ligand_activities, and ligand-to-target gene links are recovered via get_weighted_ligand_target_links. The authors of NicheNet have shown that the Pearson correlation coefficient between a ligand's target prediction and observed transcriptional response is the most informative metric of ligand activity[20]. Therefore, we select a Pearson coefficient threshold (by default 0.075) to define active ligands in each cell. Finally, we weight individual values of $V_{N_i N_j}$; $\sqrt{N_i^{l_1} * N_j^{r_1}}$ is weighted proportionally to the corresponding Pearson coefficient where the pearson coefficient of $N_i^{l_1}$ is greater than the user-defined Pearson threshold.

Downstream analysis of weighted cell-cell interaction matrices

M can be treated analogously to the gene expression matrix and used for downstream analysis tasks like dimensionality reduction. After generation and (optional) weighting of ***M*** by active ligands, ***M*** is placed into an assay of a Seurat object for downstream analysis. ***M*** is scaled by ScaleData, latent variables found via PCA, and the top principal components (identified by ElbowPlot for each dataset; default 10) used to embed the dataset in two dimensions using UMAP[48]. Neighbor graphs are constructed by FindNeighbors, which can then be clustered via modularity optimization graph-based clustering[49] as implemented by Seurat's FindClusters[47]. Differential ligand-receptor edges between clusters, cell types, or samples can be identified via FindMarkers. Scriabin provides several utility functions to facilitate visualization of gene expression profiles or other metadata on Seurat objects built from cell-cell interaction matrices.

Summarized interaction graph and binning workflow

Generation of summarized interaction graph

Because ***M*** scales exponentially with dataset size, it is frequently impractical to calculate ***M*** for all cell-cell pairs N_i,N_j. In this situation, Scriabin supports two workflows that do not require aggregation or subsampling. In the first workflow, a summarized cell-cell interaction graph ***S*** is built in lieu of ***M*** where $S_{i,j} = \Sigma V_{N_i N_j}$. ***S*** thus represents the magnitude of predicted interaction across all cognate ligand-receptor pairs expressed by all sender-receiver cell pairs. As for the cell-cell interaction matrix ***M***, ***S*** is also corrected for differences due to sequencing depth by linear regression. ***S*** may optionally be weighted by upstream regulome as described above. The second workflow is described below under "**Interaction program discovery workflow**".

Dataset binning for comparative CCC analyses

Once summarized interaction graphs are built for multiple samples, alignment of these graphs requires knowledge about which cells between samples represent a shared molecular state. The goal of binning is to assign each cell a bin identity so that ***S*** from multiple samples can be summarized into equidimensional matrices based on shared bin identities.

The binning process begins by constructing a shared nearest neighbor (SNN) graph via FindNeighbors defining connectivity between all cells to be compared. Alternate neighbor

graphs, for example those produced using Seurat's weighted nearest neighbor workflow which leverage information from multi-modal references, can also be used. Next, mutual nearest neighbors (MNNs) are identified between all sub-datasets to be compared via Seurat's integration workflow (FindIntegrationAnchors)[21]. Briefly, two sub-datasets to be compared are placed into a shared low dimensional space via diagonalized canonical correlation analysis (CCA), and the canonical correlation vectors are log-normalized. Normalized canonical correlation vectors are then used to identify k-nearest neighbors for each cell in its paired dataset and the resulting MNN pairings are scored as described[21]. Low scoring MNN pairings are then removed, as they have a higher tendency to represent incorrect cell-cell correspondences when orthogonal data is available (**Supplementary Figure 3**).

For each cell that participates in an MNN pair, Scriabin defines a bin as that cell and all cells with which it participates in an MNN pair. Next, Scriabin constructs a connectivity matrix $\boldsymbol{G}$ where $\boldsymbol{G}_{i,j}$ is the mean connectivity in the SNN graph between cell *i* and the cells within bin *j*. Each cell C_i is assigned a bin identity of the bin *j* with which it shares the highest connectivity in $\boldsymbol{G}$. Next, we optimize for the set of bins that results in the best representation of all samples. Bins with the lowest total connectivity and lowest multi-sample representation in $\boldsymbol{G}$ are iteratively removed and cell bin identities re-scored until the mean sample representation of each bin plateaus. Within-bin connectivity and sample representation are further improved by re-assigning cells that result in better sample representation of an incompletely represented bin while maintaining equal or greater SNN connectivity with the cells in that bin. Finally, incompletely represented bins are merged together based on SNN connectivity. At the end of this process, each cell will thus have a single assigned bin identity, where each bin contains cells from all samples to be compared.

Statistical analysis of bin significance

Bins are then tested for the statistical significance of their connectivity structure using a permutation test. For each bin, random bins of the same size and number of cells per sample are generated iteratively (by default 10,000 times). The connectivity vector of the real bins is tested against each of the random bins by a one-sided Mann-Whitney U test. If the bin fails 500 or more of these tests (p-value 0.05), it is considered non-significant.

Because bin SNN connectivity is generally non-zero, but randomly sampled cells generally have an SNN connectivity of zero, this strategy will tend to return most bins as statistically significantly connected. Thus, we recommend passing high-resolution cell type labels to the binning significance testing. In this situation, randomly generated bins are generated by randomly selecting cells from the same sample and cell type annotation, and the permutation test proceeds as described above. Bins where greater than a threshold (by default 95%) of cells belong to the same cell type annotation are automatically considered significant. This avoids rare cell types that may only form a single bin from being discarded. Cells that were assigned to bins which failed the significance testing are re-assigned to the bin with which they share the highest SNN connectivity.

Identification of variable bins

For each bin, a Kruskal-Wallis test is used to assess differences in the magnitude of CCC between cell-cell pairs from different samples. The Kruskal-Wallis p-value and test statistic can be used to identify which bins contain cells that exhibit the highest change in prediction interaction scores. This set of sender and receiver cells can then be used to construct ***M*** as described above.

Interaction program discovery workflow

Iterative approximation of a ligand-receptor pair topological overlap matrix (TOM)

An alternative to the summarized interaction graph workflow is to instead identify co-expressed ligand-receptor pairs, which we refer to as "interaction programs." This approach represents an adaptation of the well-established weighted gene correlation network analysis (WGCNA)[22] and is scalable to any dataset size and still permits analysis of CCC at single-cell resolution. The first step in this workflow is to generate a signed covariance matrix of ligand-receptor pairs for each sample, defined as:

$$s_{ij}^{signed} = 0.5 + 0.5cor(lr_i, lr_j),$$

where lr_i, lr_j are individual ligand-receptor pair vectors of ***M***. In large datasets, s_{ij}^{signed} is approximated by iteratively generating subsets of ***M***. s_{ij}^{signed} is next converted into an adjacency matrix via soft thresholding:

$$a_{ij} = (s_{ij}^{signed})^{\beta},$$

where β is the soft power. Soft power is a user-defined parameter that is recommended to be the lowest value that results in a scale-free topology model fit of > 0.8. Next, this adjacency matrix is converted into a TOM as described[50]. This process proceeds separately for each sample to be analyzed in a multi-sample dataset.

Identification and significance testing of interaction programs

The TOM is hierarchically clustered, and interaction programs identified through adaptive branch pruning of the hierarchical clustering dendrogram. Intramodular connectivity for each ligand-receptor pair in each interaction program is then calculated as described[51]. If interaction programs are being discovered in a multi-sample dataset, similar modules (defined by Jaccard overlap index above a user-defined threshold) are merged. Next, interaction programs are then tested for statistically significant co-expression structure via a permutation test where random interaction programs are generated 10,000 times. The correlation vector of the real module is tested against each of the random modules by a one-sided Mann-Whitney U test. If the module fails 500 or more of these tests (p-value 0.05), it is considered non-significant. Each sample is tested for significant correlation of each module.

Downstream analysis of interaction programs

Single cells are scored separately for the expression of the ligands and receptors of each significant module with Seurat's AddModuleScore. This function calculates a module score by

comparing the expression level of an individual query gene to other randomly-selected control genes expressed at similar levels to the query genes, and is therefore robust to scoring modules containing both lowly and highly expressed genes, as well as to scoring cells with different sequencing depth. Scriabin includes several utility functions to conveniently visualize interaction program expression for sender and receiver cells.

Identification of longitudinal CCC circuits

A longitudinal CCC circuit is composed of S_1-L_1-R_1-S_2-L_2-R_2, where S are sender cells and R are receiver cells at timepoints 1 and 2, and where L_1 is expressed by/sensed by S_1/R_1 and L_2 is expressed by/sensed by S_2/R_2 . For computational efficiency, construction of longitudinal CCC circuits starts at the end of the circuit and proceeds upstream. First, ligands L_2 predicted by NicheNet to be active in receiver cells at timepoint 2 are identified. Next, sender cells that express L_2 *and* have the L_2 in its per-cell gene signature are identified. Among the bins occupied by these S_2 candidates, Scriabin then searches for receiver cells at timepoint 1 that occupy the same bin and have the corresponding timepoint 2 ligand L_2 within its list of upregulated target genes and identifies the ligand(s) L_1 predicted by NicheNet to result upregulation of that target. Finally, Scriabin identifies S_1 candidates that express the timepoint 1 ligands L_1 and have L_1 in its per-cell gene signature. S_1-R_1-S_2-R_2 cell groups that meet these criteria are retained for further analysis. This process repeats for every pair of timepoints. Finally, Scriabin searches for overlap between circuits of sequential time point pairs to identify circuits that operate over more than two timepoints.

Processing, analysis, and visualization of public scRNA-seq datasets

Datasets of PBMCs (pbmc5k and pbmc10k) were downloaded from 10X genomics (https://www.10xgenomics.com/resources/datasets). scRNA-seq data of human leprosy granulomas[29] was downloaded from https://github.com/mafeiyang/leprosy_amg_network. Data from developing fetal intestine[37] was acquired from the cellxgene portal: https://cellxgene.cziscience.com/collections/60358420-6055-411d-ba4f-e8ac80682a2e. Data of longitudinal responses to SARS-CoV-2 infection in HBECs[42] was downloaded from the Gene Expression Omnibus accession #GSE166766. In each case, we acquired raw count matrices or processed Seurat objects containing raw count matrices. Any upstream processing was performed as described in the corresponding manuscripts.

The R package Seurat[21,52] was used for data scaling, transformation, clustering, dimensionality reduction, differential expression analysis, and most visualizations. Raw count matrices from Ravindra, et al.[42] required filtering before downstream analysis; cells meeting the following criteria were kept: >1,000 UMIs, <20,000 UMIs, >500 unique features, <0.85 UMI-to-unique feature ratio, <20% UMIs of mitochondrial origin, <35% reads from ribosomal protein-encoding genes. Pbmc5k and pbmc10k datasets from 10X genomics were filtered to enforce a minimum number features per cell of 200 and to remove genes not expressed in at least 3 cells. Data were scaled and transformed and variable genes identified using the SCTransform function, and linear regression performed to remove unwanted variation due to cell quality (% mitochondrial reads, % rRNA reads). PCA was performed using the 3,000 most highly variable genes. The

first 50 principal components (PCs) were used to perform UMAP to embed the dataset into two dimensions. Next, the first 50 PCs were used to construct a shared nearest neighbor graph (SNN; FindNeighbors) and this SNN used to cluster the dataset (FindClusters).

Cell type annotations were provided for the Ma, et al. and Fawkner-Corbett, et al. datasets, which were used for downstream analytical tasks. For the Ravindra, et al. dataset, manual annotation of cellular identity was performed by finding differentially expressed genes for each cluster using Seurat's implementation of the Wilcoxon rank-sum test (FindMarkers()) and comparing those markers to known cell type-specific genes listed in the Ravindra, et al.[42] PBMC datasets were annotated by weighted nearest neighbor projection and label transfer from a multi-modal PBMC reference as described[47,53].

Comparative analyses between Scriabin and published CCC analysis methods

Pbmc5k and pbmc10k datasets from 10X genomics were used to benchmark the computational efficiency of Scriabin. For single dataset analyses, pbmc5k was randomly subsetted to multiple dataset sizes. Cell type annotations were passed to Connectome[27], NATMI[17], CellChat[15], iTALK[16], and SingleCellSignalR (SCA)[26], which were run using default parameters defined by Liana[28]. The time for these methods to return results was compared to a version of Scriabin that generated and visualized a full-dataset summarized interaction graph and returned pseudobulk ligand-receptor pair scores for each cell type annotation. Connectome[27] is the only of these packages that supports a full comparative workflow. For comparative analysis, we analyzed differences in CCC between the pbmc5k and pbmc10k datasets. We compared Connectome's total runtime to the runtime of Scriabin to generate full dataset summarized interaction graphs, perform dataset binning, and visualize the most perturbed bins.

Multiple ligand-receptor resources compiled by Liana[28] were used to compare results returned by published CCC analysis methods and Scriabin. The following results parameters were used from each method: prob (CellChat), LRscore (SingleCellSignalR), weight_norm (Connectome), weight_comb (iTALK), edge_avg_expr (NATMI). To visualize the overlap in results between the methods and resources, we extracted the top 1,000 results from each method-resource pair and calculated the Jaccard index between these top results (as described by[28]).

Analysis of spatial transcriptomic datasets with Scriabin

To evaluate if Scriabin returns biologically meaningful CCC edges, we downloaded spatial coordinates and gene expression count matrices from 11 spatial transcriptomic datasets from the 10X Visium platform available at https://www.10xgenomics.com/resources/datasets. We treated each count matrix analogously to scRNA-seq data, performing data transformation and dimensionality reduction as described above. We calculated per-cell gene signatures for each dataset based on variable genes across the dataset, which we then used to rank ligands based on their predicted ability to result in the observed gene expression profile using NicheNet[20]. Next, we constructed a summarized interaction graph using a ligand-receptor pair database that was restricted to membrane-bound ligands and receptors, which we weighted according to the

predicted ligand activities. Finally, we compared the distance quantile of the top 1% of interacting cell-cell pairs compared to randomly permuted distances.

SUPPLEMENTARY INFORMATION

Supplementary Text

Conceptual requirements for dataset alignment for comparative analyses of summarized interaction graphs

Comparing summarized interaction graphs from multiple samples requires that cells from different samples share a set of labels or annotations denoting what cells represent the same identity. Each identity class to be compared then requires representation from each of the samples to be compared. This annotation typically comes in the form of coarse, low-resolution labels like cluster or cell type calls. We sought to minimize the degree of agglomeration required for comparative CCC analysis by maximizing the resolution of cell type identity labels.

We hypothesized that high-resolution clustering or sub-clustering is an inadequate solution because greater transcriptional perturbation between samples necessitates lower clustering resolutions to capture representation from each sample. To illustrate this observation, we analyzed a toy dataset of peripheral blood monocytes from a longitudinal experiment (**Supplementary Figure 3**). Cells from the week 4 timepoint show a high degree of transcriptional perturbation from the other samples, visually evidenced by little overlap in low dimensional manifold embeddings. At default clustering resolution, the cluster that constitutes this sample does not contain cells from two of the samples we wish to compare. Decreasing the cluster resolution (ie. increasing the degree of agglomeration) to 0.05 improves representation of other samples but still fails to capture cells from all timepoints in each cluster. An optimal strategy for comparing summarized interaction graphs thus involves manifold alignment rather than clustering.

Scriabin's binning workflow

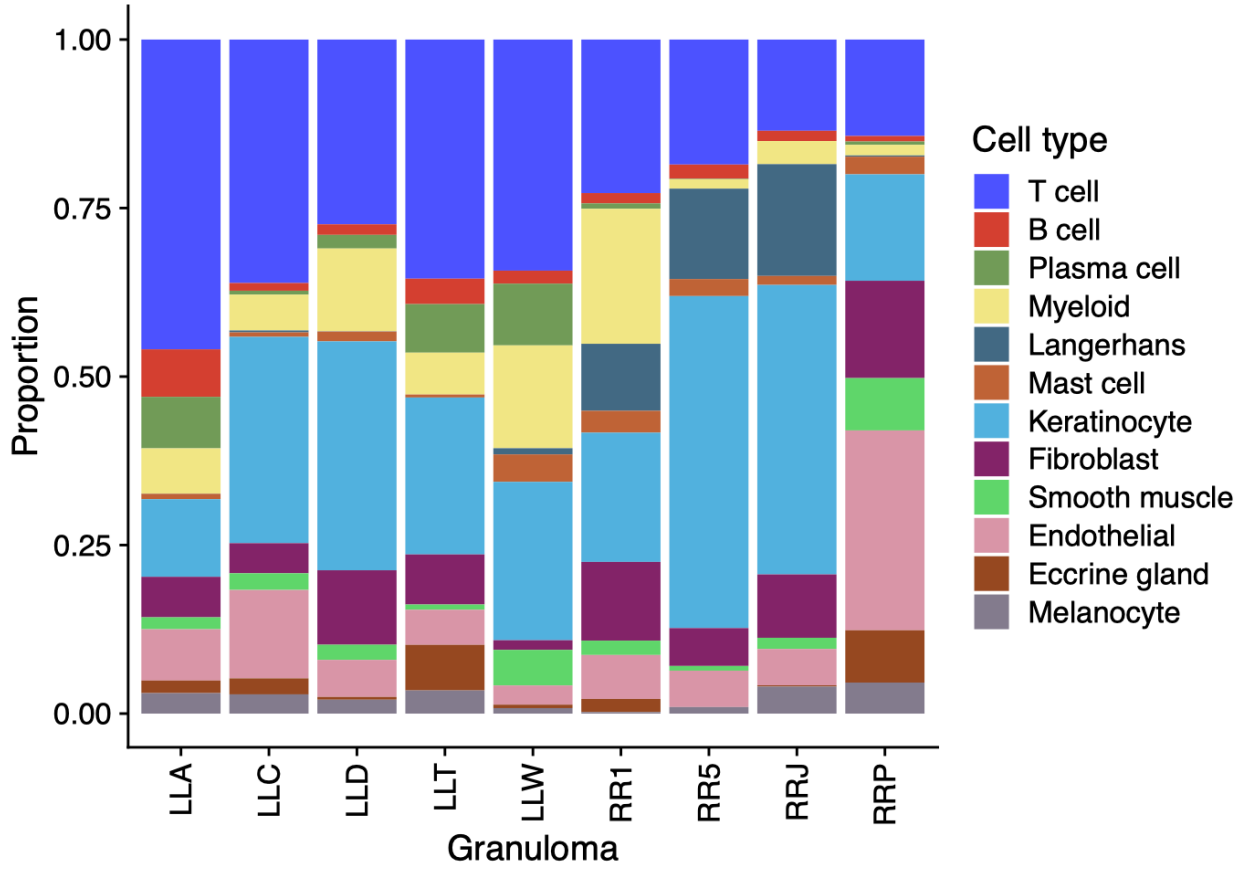
Identifying mutual nearest neighbors (MNNs)[25] between datasets has been implemented in several methods for dataset integration[21,54–56]. We reasoned that because MNNs represent pairs of cells with a shared molecular state, MNNs themselves encode single-cell resolution inter-dataset correspondences that could be generalized to high-resolution identity labels for all cells in a dataset. We refer to this process as "binning" to distinguish it from graph-based clustering. Binning with Scriabin begins by identifying MNNs between all datasets to be compared, as implemented by Seurat v3[21]. MNNs are then filtered, as we have observed that cross-cell type anchor pairs tend to have lower scores (**Supplementary Figure 4**). Cells are initially binned with the set of MNNs with which they share the highest connectivity in the shared nearest neighbor (SNN) graph, and these bin assignments are further optimized for SNN connectivity and representation of all datasets to be compared. At the end of the binning process, each cell will have a high-resolution bin identity linking it to at least one cell from all other datasets to be compared. These identities can be used as the basis for comparative analysis of CCC that maintain near single-cell resolution.

We also evaluated the performance of Scriabin's binning strategy for comparative CCC analyses. In a toy dataset of ~14,000 cells from nine sub-datasets, Scriabin identified a total of 456 bins with a median bin size of 25 cells, maintaining near single-cell resolution (**Supplementary Figure 4**). Additionally, cells from each bin generally shared the same orthogonal reference-based cell type annotation (**Supplementary Figure 4**). For example, all plasmacytoid dendritic cells (pDCs) fell into a single bin that was composed of only pDCs. Bins whose cells did not share the same cell type annotation generally shared related annotations, for instance, intermediate and naive B cells frequently occupied the same bin (**Supplementary Figure 4**).

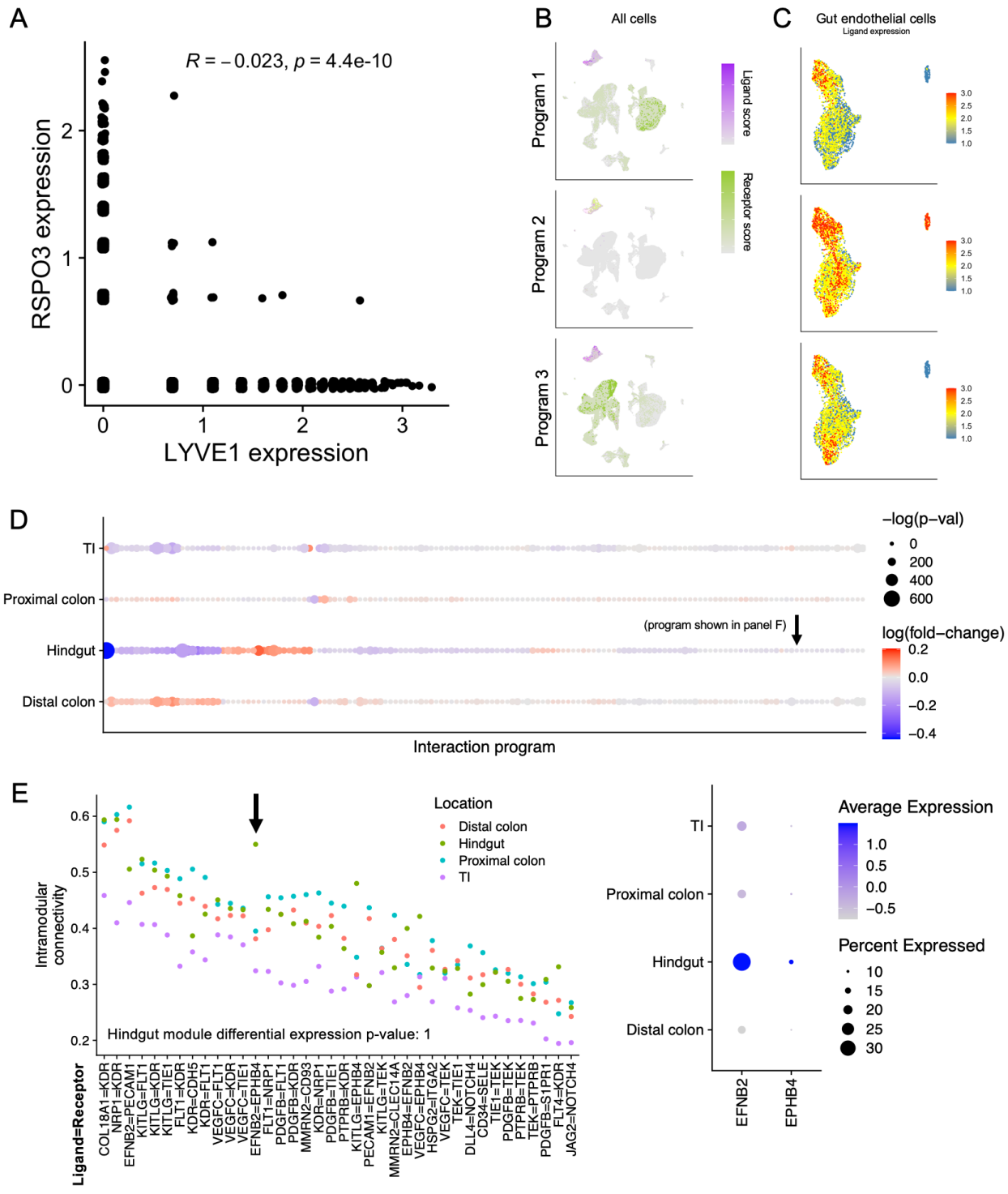
Heterogeneity of interaction program structure and expression

The discovery of interaction programs in the intestinal development dataset[37] revealed several programs that were significantly up- or down-regulated between anatomical sampling locations, reflecting either changes in the magnitude of program expression or a shift in the proportions of cell types expressing the program (**Supplementary Figure 2**). We also hypothesized that interaction program structure may vary between samples even if the expression magnitude of the program does not change. Thus, we calculated intramodular connectivity scores for each gene in each sampling location analyzed. We identified ligand-receptor pairs in non-differentially expressed interaction programs that had differential intramodular connectivity between sampling locations (**Supplementary Figure 2**). For example, the *EFNB2* - *EPHB4* ligand-receptor pair has higher intramodular connectivity in hindgut than other anatomical locations, and the hindgut is the location that expresses the highest level of both of these genes (**Supplementary Figure 2**). This indicates that, in the hindgut, when *EFNB2* and *EPHB4* are expressed, they are co-expressed along with other ligand receptor pairs in this interaction program. Collectively, this analysis provides an example of the heterogeneity, specificity, and nuanced co-expression patterns that can be revealed through the scalable discovery of interaction programs in atlas-scale datasets.

Supplementary Figures

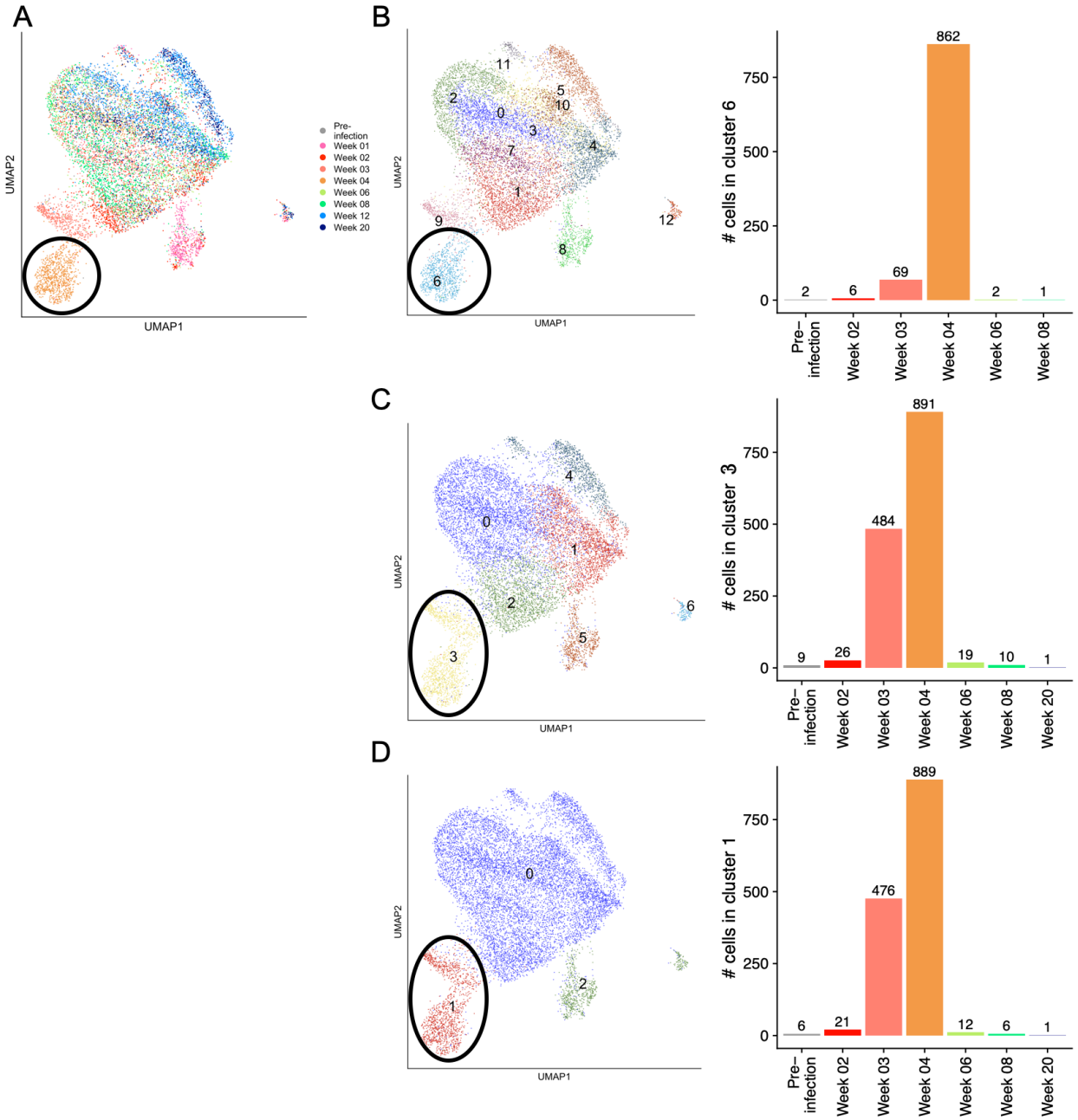


Supplementary Figure 1: **Cell type proportions in leprosy granuloma dataset.** Bar graph depicting cell proportions per granuloma in the dataset of Ma, et al[29]. Author-provided cell type annotations are used for analysis.

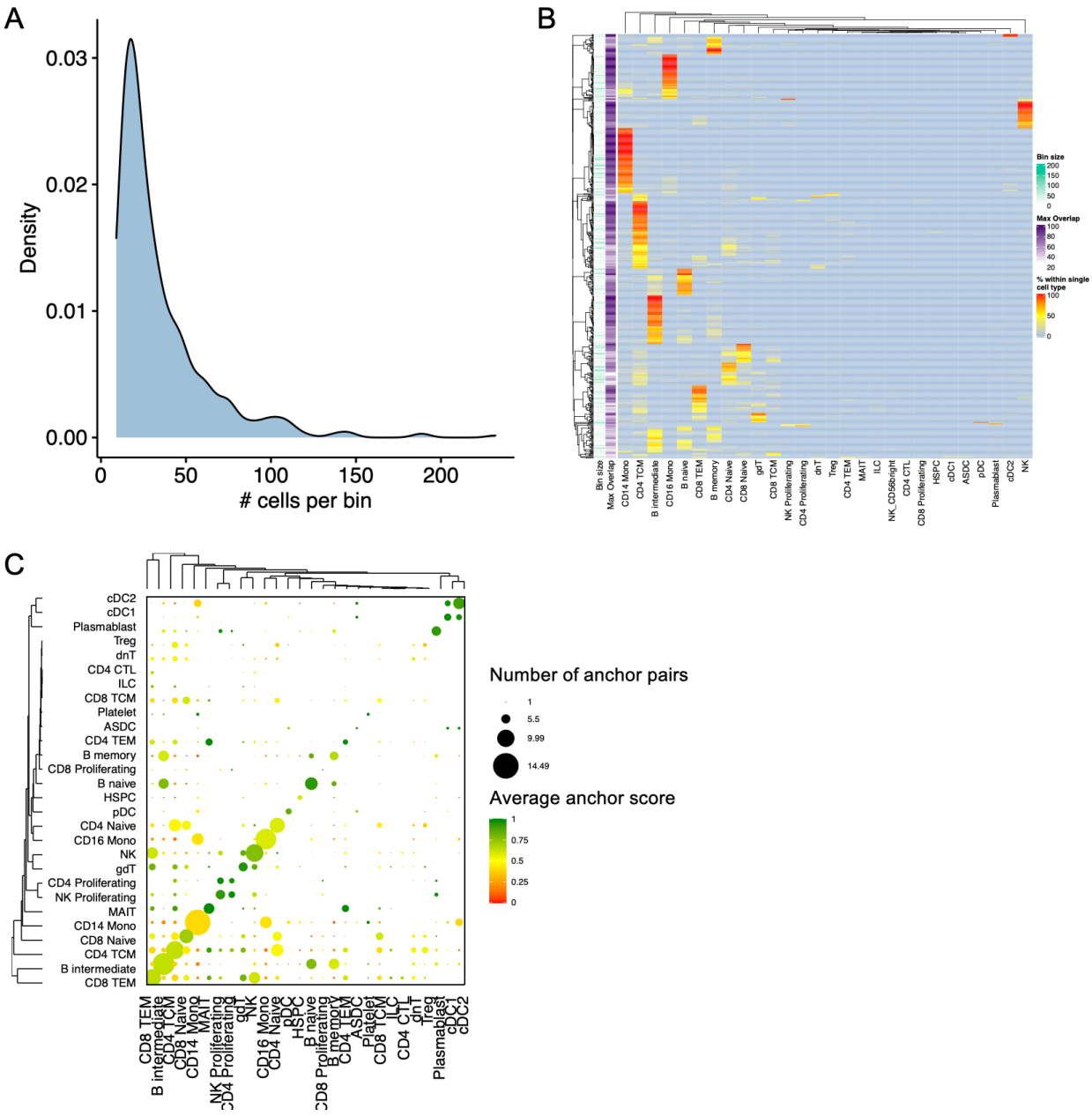


Supplementary FIgure 2: Co-expressed interaction programs in intestinal development. **A**) Scatter plot depicting expression of LEC marker *LYVE1* and *RSPO3*. **B)** UMAP projections of ligand (shades of purple) or receptor (shades of green) expression in 3 gut endothelial cell-specific modules. **C)** UMAP projection of gut endothelial cells colored by expression of ligands in the interaction programs depicted in (**B**). **D)** Dot plot depicting the expression

fold-change and Bonferroni-corrected Wilcoxon rank-sum test 2-sided p-values of interaction program expression in each anatomical location. **E)** Intramodular connectivity scores for each ligand-receptor pair in each anatomical location for the module indicated by the arrow in (**D**). The black arrow in (**E**) indicates the genes whose average and percent expression are plotted to the right.



Supplementary Figure 3: Highly perturbed samples require a higher degree of aggregation for dataset alignment. A toy dataset of peripheral blood monocytes from a longitudinal dataset was analyzed. **A)** UMAP projection colored by time point. **B-D)** UMAP projections (left) colored by cluster identity, and bar plot depicting per timepoint cluster membership in the cluster principally occupied by sample Week 04 (right). Cluster resolutions: 1 (default, **B**), 0.3 (**C**), 0.05 (**D**).



Supplementary Figure 4. Robustness analysis of Scriabin's binning workflow. A toy dataset of ~14,000 peripheral blood mononuclear cells (PBMCs) from nine sub-datasets was analyzed. **A)** Density plot depicting the number of cells in each bin. The median bin size in this analysis is 25 cells. **B)** Heatmap depicting overlap between bin identity and cell type annotations. Each row sums to 100%. An SNN graph was used to assess cell-cell connectivity for the binning workflow. Cell type annotations are transferred from a reference dataset and are thus orthogonal to the data used to generate the bins. **C)** Dot plot depicting the cell type annotations and scores for the anchor pairs used to generate the bins depicted in (**B**).

REFERENCES

1. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell–cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology* **26**, 12–23 (2021).
2. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2020).
3. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
4. Yosef, N. & Regev, A. Writ large: Genomic dissection of the effect of cellular environment on immune response. *Science* (2016) doi:10.1126/science.aaf5453.
5. Ramilowski, J. A. *et al.* A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
6. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
7. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
8. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611–1624.e24 (2017).
9. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
10. Pavličev, M. *et al.* Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome Res.* **27**, 349–361 (2017).
11. Zepp, J. A. *et al.* Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung. *Cell* **170**, 1134–1148.e10 (2017).
12. Cohen, M. *et al.* Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in

Macrophage Imprinting. *Cell* **175**, 1031–1044.e18 (2018).

13. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
14. Raredon, M. S. B. *et al.* Single-cell connectomic analysis of adult mammalian lungs. *Sci Adv* **5**, eaaw3851 (2019).
15. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1–20 (2021).
16. Wang, Y. *et al.* iTALK: an R Package to Characterize and Illustrate Intercellular Communication. doi:10.1101/507871.
17. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).
18. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
19. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, e9923 (2021).
20. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
22. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
23. Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.* **39**, 1095–1102 (2021).
24. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is

inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

25. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
26. Cabello-Aguilar, S. *et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* **48**, e55–e55 (2020).
27. Raredon, M. S. B. *et al.* Connectome: computation and visualization of cell-cell signaling topologies in single-cell systems data. *bioRxiv* 2021.01.21.427529 (2021) doi:10.1101/2021.01.21.427529.
28. Dimitrov, D. *et al.* Comparison of Resources and Methods to infer Cell-Cell Communication from Single-cell RNA Data. *bioRxiv* 2021.05.21.445160 (2021) doi:10.1101/2021.05.21.445160.
29. Ma, F. *et al.* The cellular architecture of the antimicrobial response network in human leprosy granulomas. *Nat. Immunol.* **22**, 839–850 (2021).
30. Gordon, S. Alternative activation of macrophages. *Nat. Rev. Immunol.* **3**, 23–35 (2003).
31. Ridley, D. S. & Jopling, W. H. Classification of leprosy according to immunity. A five-group system. *Int. J. Lepr. Other Mycobact. Dis.* **34**, 255–273 (1966).
32. Flynn, J. L. *et al.* An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection. *J. Exp. Med.* **178**, 2249–2254 (1993).
33. Herbst, S., Schaible, U. E. & Schneider, B. E. Interferon gamma activated macrophages kill mycobacteria by nitric oxide induced apoptosis. *PLoS One* **6**, e19105 (2011).
34. Ní Cheallaigh, C. *et al.* A Common Variant in the Adaptor Mal Regulates Interferon Gamma Signaling. *Immunity* **44**, 368–379 (2016).
35. Verhagen, C. E. *et al.* Reversal reaction in borderline leprosy is associated with a polarized

shift to type 1-like Mycobacterium leprae T cell reactivity in lesional skin: a follow-up study. *J. Immunol.* **159**, 4474–4483 (1997).

36. Teles, R. M. B. *et al.* Identification of a systemic interferon-γ inducible antimicrobial gene signature in leprosy patients undergoing reversal reaction. *PLoS Negl. Trop. Dis.* **13**, e0007764 (2019).
37. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826.e23 (2021).
38. Biton, M. *et al.* T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. *Cell* **175**, 1307–1320.e22 (2018).
39. Goto, N., Imada, S., Deshpande, V. & Yilmaz, Ö. H. Lymphatics constitute a novel component of the intestinal stem cell niche. *bioRxiv* 2022.01.28.478205 (2022) doi:10.1101/2022.01.28.478205.
40. Niec, R. E. *et al.* A lymphatic-stem cell interactome regulates intestinal stem cell activity. *bioRxiv* (2022).
41. Darling, T. K. & Lamb, T. J. Emerging Roles for Eph Receptors and Ephrin Ligands in Immunity. *Front. Immunol.* **10**, 1473 (2019).
42. Ravindra, N. G. *et al.* Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol.* **19**, e3001143 (2021).
43. Gressner, O. A., Peredniene, I. & Gressner, A. M. Connective tissue growth factor reacts as an IL-6/STAT3-regulated hepatic negative acute phase protein. *World J. Gastroenterol.* **17**, 151–163 (2011).
44. Sack, G. H., Jr. Serum amyloid A - a review. *Mol. Med.* **24**, 46 (2018).
45. Xu, J. *et al.* SARS-CoV-2 induces transcriptional signatures in human lung epithelial cells

that promote lung fibrosis. *Respir. Res.* **21**, 182 (2020).

46. Andreatta, M. *et al.* Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat. Commun.* **12**, 2965 (2021).
47. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021) doi:10.1016/j.cell.2021.04.048.
48. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.
49. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 471 (2013).
50. Yip, A. M. & Horvath, S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* **8**, 22 (2007).
51. Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC Syst. Biol.* **1**, 24 (2007).
52. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
53. Wilk, A. J. *et al.* Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19. *J. Exp. Med.* **218**, (2021).
54. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
55. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
56. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).