CINS: Cell Interaction Network inference from Single cell expression data

Ye Yuan[1,2], Carlos Cosme Jr.[3], Taylor Sterling Adams[3], Jonas Schupp[3], Koji Sakamoto[3], Nikos Xylourgidis[3], Matthew Ruffalo[4], Jiachen Li[1], Naftali Kaminski[3], Ziv Bar-Joseph[2,4*].

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China

[2]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[3]Section of Pulmonary, Critical Care and Sleep Medicine, Yale University School of Medicine, New Haven, CT 06520, USA.

[4]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

*e-mail: zivbj@cs.cmu.edu

Abstract

Studies comparing single cell RNA-Seq (scRNA-Seq) data between conditions mainly focus on differences in the proportion of cell types or on differentially expressed genes. In many cases these differences are driven by changes in cell interactions which are challenging to infer without spatial information. To determine cell-cell interactions that differ between conditions we developed the Cell Interaction Network Inference (CINS) pipeline. CINS combines Bayesian network analysis with regression-based modeling to identify differential cell type interactions and the proteins that underlie them. We tested CINS on a disease case control and on an aging mouse dataset. In both cases CINS correctly identifies cell type interactions and the ligands involved in these interactions improving on prior methods suggested for cell interaction predictions. We performed additional mouse aging scRNA-Seq experiments which further support the interactions identified by CINS.

Introduction

The ability to profile the expression of genes at the single cell level has revolutionized gene expression studies. Single cell RNA-Seq (scRNA-Seq) studies resulted in insights related to the cell type composition of tissues (1,2), changes in cell type composition in various diseases and states (3), various differentiation pathways used within cells (4) and more. However, while scRNA-Seq provides valuable information about expression within cells, it is hard to use it to study interaction between cells. The main problem is that once cells are extracted it is very challenging to determine the spatial relationships among them (5).

A number of methods have been introduced recently to identify ligand receptor interactions in scRNA-Seq studies (6,7). While these methods differ in the exact formulation and statical analysis, they all focus on finding correlations between ligands expressed in one cluster (or cell type) and receptors expressed in another. This works well for studies that are analyzing a single condition (for example expression in a specific tissue or at a specific time point) but does not fully utilize information in case-control studies single cell studies (8,9). Unlike single condition studies, in addition to differences in expression case-control studies also provide information on differences in the *proportions* of different cell types between the conditions. Such information can be very useful in determining which cell types interact. When cell type proportions are correlated between two conditions (for example both high in one and low in the other) it may indicate that they are likely to interact (10,11). As we show, this information greatly improves the ability to correctly infer cell-cell interactions from scRNA-Seq data.

In addition to methods that attempt to infer cell-cell interaction information from scRNA-Seq, a number of technologies have emerged for spatially profiling single cell expression data (12-15). These technologies often combine Fluorescence in situ hybridization (FISH) with rapid sequencing

to provide information on the spatial expression of thousands of genes at various resolutions (16,17). A number of recent computational methods have been developed to allow for the study of signaling pathways involved in cell-cell interactions from this type of spatially-resolved expression data (18). However, while spatial transcriptomics studies are promising there are several challenges involved in employing them to study intercellular interactions. First, current commercial spatial transcriptomics platforms do not profile cells at the single cell level. Most labs do not have access or ability to perform such studies at the single cell resolution. More importantly, spatial transcriptomics often requires the fixation of the samples which limits their usage and can negatively impact their ability to accurately profile molecular quantities (16). Finally, spatial transcriptomics methods can scan only a small region of the tissue and so cannot be applied to large number of conditions and samples that are studied using scRNA-Seq.

Here we present a new method, the Cell Interaction Network Inference (CINS) pipeline, that infers cell type interactions in case control scRNA-Seq studies. CINS involves two major steps. First, it uses scRNA-Seq data from multiple samples of a similar condition (i.e. disease, age, etc.) to learn Bayesian networks which highlight the cell types whose distributions are co-varying under different conditions. Next, for the high scoring differential interactions identified in the Bayesian network analysis, CINS learns a regression model with ligand-target interaction matrix (6) that identifies the key ligands and targets that participate in the interactions between these cell types.

We tested CINS by applying it to both, disease and aging datasets. We show that CINS correctly identifies known interacting cell type pairs and ligands associated with these interactions and improves upon prior methods for inferring ligand-receptor interactions in scRNA-Seq data. We also discuss several novel predictions made by CINS. Finally, we show that a number of CINS predicted cell type interactions are supported by a new scRNA-Seq lung aging dataset we profiled.

Results

The Cell Interaction Network Inference (CINS) Pipeline

We developed the Cell Interaction Network Inference (CINS) pipeline which uses single cell (sc) RNA-seq expression data to infer cell-cell interactions (**Fig. 1**). Given repeated experiments of the same condition / system CINS uses annotated cell type information to construct a Bayesian network (BN) that models causal relationships between different cell types. For this, CINS first discretizes the proportion data for each cell type using a Gaussian Mixture Model (GMM) with only two components and then learns a BN that models the joint probability distribution of the cell type mixtures observed for each sample. High scoring differential causal relationships are determined based on bootstrapping. Next, for each of the high scoring differential pairs identified we infer the genes involved in the interactions by learning a ligand-target regression (LTR) model with ligand-target interaction database from NicheNet (6). The LTR model aims to explain changes in target genes as a function of changes in their activating ligands allowing CINS to identify the most significant ligands that regulate the cell-cell interactions.

Inferring cell type interactions using Bootstrapped Bayesian Network

We first studied a lung disease scRNA-Seq dataset (8). The lung disease dataset contained scRNA-Seq data for 28 healthy (controls) and 32 Idiopathic Pulmonary Fibrosis (IPF) individuals. A total of 250,942 cells were profiled for these individuals. Cell type annotations were assigned based on the original study and we used the detailed assignments that provided information on 39 cell types.

We used CINS to explore differential cell type interactions between IPF and control samples. For this, we constructed two different networks based on the cells profiled for each condition. We next performed bootstrap analysis to determine the score of each edge in each condition. Edges that

appear in the majority of bootstrap iterations likely represent real relationships in the data rather than noise (19,20). Resulting BNs for the two conditions are presented in **Fig. 2A&B.** As the figures show, there are some edges that appear for both conditions. These include Basal to Goblet cell interactions, which agrees with the fact that club cell's attachment sites are provided by Basal cell (21). However, there are also many differences between edges selected for the two condition networks. **Tab. 1** summarized the top differences based on the signed difference in edge count in 100 bootstrap iterations for IPF and control (See **Tab. S1** for differences for all detected edges). Several of the highest scoring edges are supported by prior work. For example, the edge from Treg to Fibroblast cell is supported by a previous study suggesting that Treg's can negatively regulate fibroblast activity (22). The edge between cDC2 and cDC1 is also supported by recent work showing that cDC2 and cDC1 are cross-talking with each other (23). Several other top scoring edges are supported by the literature as referenced in **Tab. 1.** We next compared the interactions predicted by CINS to interactions predicted by CellPhoneDB, iTALK, and NicheNet (Methods), which are all popular methods for inferring ligand-receptor based cell interactions (6,7,24). As can be seen, in **Tab. S2**. unlike CINS which identified a diverse set of cell type interactions, almost all interactions predicted by CellPhoneDB involved Goblet cells (18 of the top 20). While there is some support for Goblet involvement in IPF (25) they only explain a small fraction (estimated to be less than 20%) of individuals with the disease and it is unlikely that they interact with almost all other cell types. Similarly, for NicheNet, almost all interactions predicted involved a single cell type, Pericyte cells. iTALK performed better, but it has only detected interactions between immune cells in the IPF lung dataset. While these are indeed of interest, the more interesting interactions are those between immune cells and fibroblast cells in the (injured) lung and none of these were identified by iTALK. In contrast, by looking at the overall distribution of cell types CINS was able

to find a more general and, as we showed, accurate set of interactions between cell types that are likely relevant for the disease.

Inferring ligand-target interactions for high scoring differential cell type pairs

While the BNs discussed above identify pairs of cell types that likely interact in disease, the network does not show which genes and protein products participate in the interactions. To infer such gene-gene interactions across cells we developed a ligand-target regression (LTR) model. For cell type pairs identified in the BNs our LTR model uses a set of ligands in the first cell type to predict the expression values of their known targets in the second cell type. The LTR model uses the LASSO algorithm which enables the identification of a small set of key ligands predicted to participate in the interaction observed in the BN. We trained the model using a five-fold cross validation strategy. See **Methods** for details.

The LTR method was applied to all high scoring differential pairs identified by the BN. **Tab. S3** presents top scoring ligands for several cell type pairs. **Tab. S4** presents top scoring ligands for one cell type pair (Fibroblast -> Lymphatic cell). Several of the top LTR ligands are known to play an important role in the activated cell (Lymphatic cell). For example, the highest scoring ligand identified by LTR is “FGF2” which was identified as a critical gene for lymphangiogenesis (26). Another highly ranked ligand, “TGFB1”, can also accelerate lymphatic regeneration in wound repair (27). **Tab. S5** presents top ranked ligands for another pair (Treg cell -> Fibroblast), several of which have also been shown to participate in the interaction between these cell types. For example, fibroblast express IL13 receptor and may behave as an inflammatory cell if stimulated by IL-13 (28), and TGFB1-3 (including TGFB1 and TGFB2 in the table) are all involved in promoting collagen production in fibroblasts (29).

Identified ligands are primarily involved in cell-cell interactions

To test if the predicted ligands are indeed impacting cell type-cell type interactions or mainly represent autocrine relationships we compared the activity of top predicted ligands within and between cell types. For this, we compared the performance of the LTR method for top edges to the performance of a similar method that only uses information from a single cell type. Specifically, if the BN predicted a high scoring differential interaction between cell types A -> B, we first trained LTR using the ligands of A and the targets of B (as we did above) and compared the performance to a LTR model which uses the ligands expressed in B to predict targets in B (autocrine model).

Results for the high scoring differential edges in the IPF and control datasets is presented in **Fig. 3A. Fig. 3B.** presents the results for the same pairs (so x axis is fixed based on the BN score) but with the LTR trained using only the ligands of the second cell type. As can be seen, when using the ligand of the predicted interacting cell type LTR obtained a higher average correlation with a p-value of 0.034 (using the scipy function in Python for computing Pearson correlation p-values). In contrast, when using the same cell type for both ligands and targets the Pearson correlation is lower (**Fig. 3B)**. We also evaluated the performance of the LTR method on the predicted cell type interactions by comparing the results we obtained with the real ligand-target interaction matrix to results obtained using a random ligand-target interaction matrix. We found that for most of the random assignments the resulting LASSO models contained only a Bias term with all coefficients set to 0 (**Fig. S3**). This indicates that expression of the ligands did not provide any useful information about the expression of the targets when using the random interaction matrix.

Application to a scRNA-Seq dataset on lung aging

We next applied CINS to another, smaller, scRNA-Seq dataset which studied lung aging in mice (9). The dataset profiled lung cells in 15 mice, 8 young (three-month, 3M) and 7 old (24-month, 24M). The 14,813 cells profiled in this study were assigned to one of 34 cell types in the original paper. We again learned 100 bootstrapped BNs for the two conditions (young and old) and compared the resulting networks. We found 11 edges to be differentially present between the two conditions when using an edge threshold count of 20 (**Fig. 4** and **Tab. S6**). These included an edge between Capillary-endothelial-cell and Type 1-pneumocyte cells which are known to jointly form thin air-blood barriers used for gas exchange (30). Another pair was Ciliated and Club cells, of which the ratio is reported to alert significantly between young and old mouse lung (9). We next performed LTR analysis on the high scoring differential edges. The top ranked ligand in Ciliated cells, TNF is known to regulate CC16 gene production, which plays a role in immunomodulatory activity in Club cells (31). Apoe, a ligand identified for the macrophage to goblet edge, is produced by macrophages to negatively modulate goblet cell hyperplasia (32).

As we did for the IPF study we compared the performance of the LTR method using ligands from the BN identified edges (A -> B) and ligands from the same cell type (B) to predict target expression for genes in B. We observed a Pearson correlation of 0.67 when using the ligands from the BN identified edges (A->B) vs. Pearson correlation of 0.31 when using the ligands from B (**Fig. 5**). And it is noticed that when randomizing the interactions the LTR method again failed to identify any significant correlation between predicted and real expression for the targets (**Fig. S3**).

Computational validation of high scoring differential edges using a second aging mouse lung dataset

To test the predictions of the aging BN and to validate them using an independent cohort we next performed additional scRNA-Seq experiments on young and old mice to generate a pilot scRNA-

Seq dataset on lung aging. For this, we profiled four young and four old mice of the *Fendrr-floxed* genotype recently generated in the Kaminski laboratory. We obtained 71,562 cells that were clustered, annotated, and assigned to 20 cell types that overlapped with the cell types assigned by Angelidis I et al. (9). The problem with both aging datasets is their small size 15 and 8 compared to 60 in IPF dataset). We could not obtain significant results using the 8 dataset aging data given its small size. Thus, we could not use it as a standalone dataset to validate the results of the larger (15 samples) datasets. Instead, we looked at the impact of combining the two. We next used the combined data (from (30) and from our new experiments) to learn a joint BN. Several of the predicted interactions were further supported by our new data. Specifically, we found 19 cell type pairs for which the addition of our new data enhanced both the presence of the edge and the direction predicted when performing the bootstrap analysis. **Tab. S8** presents the top 10 enhanced pairs based on the overall bootstrap score (See **Tab. S11** for all enhanced pairs). For example, the interaction between Neutrophils and Gamma Delta T cell is enhanced from edge count of 40 to 61 and was reported by recent studies that neutrophils can suppress Gamma Delta T cell's activation involved in the resolution of inflammation (33). And the interaction between B Cell and CD4+ T Cell is enhanced from -16 to -19 (being negative means that old lung has less), and is supported by other studies that B cell will activate CD4 T cells in human cutaneous leishmaniasis infection led by Viannia (34). In addition, we also found that T-cell-B-cell interactions were calculated to occur less often in older samples, which further validates the comparison between old and young mice (35).

We next focused on the top five predicted interactions in **Tab. S8** (all with an absolute enhanced bootstrap score larger than 15). Permutation analysis indicates that identifying such a large number of edges supported by both studies is significant (p-value = 0.05, **Methods** and **Fig. 6,** and see

Tab. S13 for result of other threshold values**)**. Specifically, we permutated the cell type fraction of the aging dataset with 8 samples, and then did the BN analysis for 1,000 times. We next calculated the fraction of enhanced pairs with certain edge threshold over the whole pairs reported. We applied LTR to the cell type pairs in **Tab. S8** to find important ligand genes. **Tab. S9** presents the top predicted ligand genes. Several of these (red font) are supported by prior studies on the interaction between these cell types. Comparisons to CellPhoneDB, iTALK and NicheNet indicated that, similar to what we observed for the IPF data, the predicted interactions are very different compared to CINS (**Fig. S4**). In addition, unlike CINS for which the overlap between the pairs identified with and without the new datasets were significant, for CellPhoneDB we did not observe significant overlap between predicted interaction pairs (**Tab. S13)**.

Discussion

To enable the study of cell type – cell type interactions using scRNA-Seq data we developed a method termed Cell Interaction Network Inference (CINS). CINS first learns a Bayesian network between cell types (BN) using repeated samples. High scoring differential cell type pairs identified by the BN are further studied to infer the ligands that regulate these interactions. CINS is implemented in python and R and can be downloaded from https://github.com/xiaoyeye/CINS.

While CINS can be applied to any dataset with multiple samples, it is most appropriate for datasets containing case and control or multiple conditions. For such datasets CINS can infer not only the high scoring differential interactions within a condition but also those interactions that differ between the condition and that may partially explain the differences between the conditions studied. Most current cell interaction tools focus on ligand gene expression information, while CINS can

make use of cell proportion as additional information and the two can prove each other mutually, further confirming the findings. The discretization of cell proportion can fit the data very well and makes it easier for BN to learn the correct structure of the network which is the major focus of CINS.

We first applied CINS to study a case and control dataset profiling lung expression from IPF patients and controls. CINS identified several differences between the interactions observed for IPF patients and for healthy individuals. These include the interaction from Treg to Fibroblast cells which is supported by a recent study that found Treg can negatively regulate fibroblast activity (22) , and the edge between cDC2 and cDC1 is also supported by recent work showing that cDC2 and cDC1 are cross-talking with each other (23).

For many of the identified high scoring differential interactions CINS was also able to identify key ligands involved in the interactions. For example, “FGF2” which was identified as a critical gene for lymphangiogenesis (26), and one more highly ranked ligand, “TGFB1”, can also accelerate lymphatic regeneration in wound repair (27).

We next applied CINS to a lung scRNA-Seq aging dataset and identified a number of high scoring differential pairs that differ between young and old mice. To validate predicted interactions we performed additional experiments in which we profiled scRNA-Seq expression in 4 additional young and old mice and then used the combined dataset to learn a joint network. As we showed, the network we learned identified a significant number of interactions that are supported by both datasets. These include the interactions between Neutrophils and Gamma Delta T cell (33), and between B Cell and CD4+ T Cell (34,35) which are both supported by previous studies. CINS was again able to identify key ligands involved in these interactions, TNF, identified as the top ligand in the interaction between neutrophils and Gamma Delta T cells was previously identified as

expressed in neutrophils (36) and as a regulator of immune cells Gamma Delta T cells (37), and TNFSF18 identified in interactions between CD4+ T cells and Vascular Endothelial Cells, was also previously reported to mediate the interactions between immune cells and endothelial cells (38).

While CINS can be successfully applied to several scRNA-Seq studies, it does have several limitations. First, it can only be applied if multiple samples are profiled since the BN part requires several repeated samples to compute relationships between cells. In addition, because BNs do not allow self edges, interactions between cells of the same type cannot be identified by CINS. Finally, since it uses a bootstrap approach to infer edge score it can miss important interactions if not enough samples and / or cells are available.

CINS is one of the first methods to enable the inference of cell type interactions in scRNA-Seq data from *repeated* samples. Given the growing popularity of this method, and its increased use in clinical studies which are currently less amenable to spatial transcriptomics techniques we believe that CINS provides a solution to an important problem that is not currently addressed.

Materials and Methods

We developed a pipeline for modeling interactions between cells of different types from scRNA-Seq data. Our method first identifies cell types that are likely interacting and then tries to provide a mechanistic model to explain how such interactions are manifested at the molecular level.

Datasets

We tested CINS using three scRNA-Seq datasets. The first compared gene expression in lungs of healthy and Idiopathic Pulmonary Fibrosis (IPF) with accession number of GSE136831 (8). This dataset contained 28 controls and 32 IPF patients with a total of 243,472 cells and the expression

levels for 45,947 genes in each cell. We used the original annotations and included in the model all 39 cell types with at least 100 cells. The second dataset studied lung aging in mice with accession number of GSE124872 (9). This dataset contained 8 three-month-old mice and 7 24-month-old mice for which a total of 14,813 cells were profiled. For each cell the expression levels of 21,969 genes were provided. Each cell was assigned by the authors to one of 34 cell types. The third dataset was a new dataset in which we profiled single cell expression in four young (25 weeks) and four old (2x 103 weeks; 2x 120 weeks, Supporting Methods) *Fendrr-floxed* mouse lungs. This dataset contained a total of 71,562 cells with expression values for 45,947 genes. These cells were originally assigned to 37 cell types based on the expression of canonical cell type markers. To combine the two aging datasets we did the following. We first normalized the gene expression data using the same method for both datasets. Next, we manually assigned a common set of cell types to both datasets so all cell type match between the two. Specifically, we identified a joint subset of 20 cell types identified by both and only used cells assigned to these cell types in our combined BN analysis (see **Tab. S10** for cell type information details).Information about ligands and their targets were obtained from a recent paper (6) which provided targets for 688 ligands.

Single-cell sequencing of *Fendrr-floxed* Mice

Animal procedures had been approved by the Institutional Animal Care and Use Committee (IACUC). We created a floxed allele of *Fendrr* via two-guide, two-oligo CRISPR/Cas mediated cleavage and recombination essentially as described in Yang et al. (19). A generated mouse which had the expected conditional allele was bred with C57BL/6J mice to establish the colony and to sort the floxed allele from any other possible mutant alleles. Three female and five male mice in two age groups (young: 23 weeks, old: ranging from 103 to 120 weeks; four mice per group) were

euthanized, and lungs were harvested and minced in small pieces with a scalpel. Lung pieces were dissociated using the enzyme Liberase TL (Roche).

Single RNA molecules of single cells were barcoded using the 10× chromium single-cell technology according to the manufacturer's instructions (Single Cell 3′ Reagent Kits v2, 10× Genomics, USA). Barcodes were used to assign reads to cells and quality control was performed to remove low quality cells (Supporting Methods). Generated sequencing data is available at GEO accession number GSE165638. A modified version of the standard Seurat pipeline was employed to normalize, cluster and annotate the raw counts single-cell expression data for downstream analysis (20). Briefly, the percent of mitochondrially-expressed genes was calculated for each individual cellbarcode using the *PercentageFeatureSet* function. Next, unique molecular identifier (UMI) counts were log normalized with a scale factor of 10,000 UMIs per cell and then natural log transformed using a pseudocount of one. Following log normalization, the top 3500 variable genes within the dataset were determined using Seurat's implementation of the *FindVariableFeatures* function with the "vst" parameter. Next, the gene-level scaling of the data was performed using the *ScaleData* function. Each feature was centered to have a mean of zero and scaled by the standard deviation of each feature. The percent of mitochondrially-expressed genes captured within each cell were regressed out during scaling by using the "vars.to.regress" parameter. To reduce the dimensionality of the dataset and to identify genes contributing the most variability to the underlying manifold of the dataset, Principal Component Analysis (PCA) was performed using the scaled data and the 3500 variable genes calculated determined for the dataset. Following exploration of the PCs (Supporting Methods), the first 75 PCs were selected for clustering and following Uniform Manifold Approximation and Projection (UMAP), a

dimensionality reduction method. The quality of subject and age representation within each cluster was assessed prior to cell type annotation to note any subject- or age-specific biases.

Cell type assignment of *Fendrr-floxed* mice

To assign a specific cellular identity to each cluster, differentially expressed markers were determined and assessed within the context of canonical marker genes. Briefly, a differential gene expression test using Wilcoxon Rank Sum test was performed that compared the gene expression within a specific cluster to expression within all cells outside of that cluster. The resulting list of cluster-specific marker genes was assessed and cell types were ascribed based on expression of canonical marker genes. Clusters displaying canonical markers for multiple cell types were flagged as multiplets and were omitted from downstream analysis.

Cell type quantification and discretization

We use the cell type annotation information provided by each study. To use Bayesian network to learn relationships between cell type we first discretize the proportion of each cell type in each sample. Discretization is cell type specific (i.e. different cell type will be assigned different values for the same proportion quantity) and is learned using an unsupervised method based on Gaussian Mixture Model (GMM) with two components. Specifically, let $\left[x_1^i, x_2^i, \cdots x_n^i \cdots, x_N^i\right]$be the fraction (percentage) of the *ith* cell type in the N samples. We learn a two components GMM for these values and then assign each value to the class with the higher likelihood for this value. The target function of the GMM aims to maximize the log likelihood:

$$l^i\left(\pi^i,\ \mu^i, \Sigma^i\right) = \sum_{n=1}^{N} \log\left(\sum_{k=0}^{1} \pi_k^i \mathcal{N}\left(x_n^i, \mu_k^i, \sigma_k^i\right)\right) \quad (1)$$

Where $\mathcal{N}$ represents gaussian distribution and ($\pi_k^i, \mu_k^i, \sigma_k^i$) represent proportion, mean and standard deviation parameters for the *kth* component of the *ith* cell type.

Following convergence, each proportion value x_n^i is assigned to one of the two classes. We assign labels to the two classes such that the component with lower mean parameter is assigned a value of 0 and the second is assigned a value of 1. This leads to a learned cell type specific cutoff such that all samples with a value less than that cutoff are assigned to 0 and all those above are assigned to 1. However, the number of 0's and 1's is not pre-determined and may be highly skewed in either direction based on the distribution of the fractions. See **Fig. S1** for examples of assignments. To learn GMMs we used the Python package "sklearn" with a maximum iteration number of 500 and a convergence threshold of 10**-4.

Learning a cell type Bayesian network

We use the discretized cell type values to learn a cell type Bayesian network. Bayesian network is a probabilistic graphical model that uses directed acyclic graph to represent joint probability distributions. The absence of an edge can indicate independence and / or conditional independence. Bayesian networks are parameterized as <*G, P*> where G = <*V, E*> is a directed acyclic graph with *V* as variables and E as directed edges, and P is the global joint distribution for all nodes *V*. Given the graph structure this probability can be decomposed into local distribution for each node, V_q, conditioned on its parent nodes as follows:

$$P(V|\Theta, G) = P(V_1, V_2, \ldots, V_Q|\Theta, G) = \prod_{q=1}^{Q} P(V_q|Pa(V_q|\Theta_q)) \quad (2)$$

Where $Pa(V_q|\Theta_q)$ is parent node set of V_q according to *G*.

To learn a Bayesian network using the discretized cell type proportion data, we iterate between network learning and parameter estimation. We initialize the network using the *Hiton Parents and Children* strategy which is based on marginal association among variables (21). Next we iterate a search strategy, that uses penalized Hill-Climbing to add, flip or remove edges based on the Bayesian Information Criterion (BIC) score when using dataset *D*, where each sample in *D* contains values for all the variables of *V*:

$$BIC = logP(D|\Theta, G) - \frac{1}{2}Dim(G)logN \quad (3)$$

where *N* is the number of samples and $Dim(G)$ is the number of parameters in the model. For this, we used the "rsmax2" function from the R library "bnlearn", which implements the iterative Penalized Maximization algorithm to construct a Bayesian network.

To obtain confidence values for edges (predicted interactions) in the network we followed previous learning methods that utilized a bootstrap strategy (22-24). For each iteration of the bootstrap we first randomly sample 80% of all single cells in the dataset. Next, we used these cells to determine cell type frequencies in each sample and to perform the discretization and network learning as described above. This step is repeated 100 times, and for which we counted the presence of all directed edges. While the direction of an edge in a Bayesian network does not always imply casual interactions (25), we observed that high scoring differential edges were also very consistent in their direction (**Tab. S1**).

Ligand-Target Regression (LTR) Model

The bootstrapping method presented above provides a small set of high scoring differential interactions between some of the cell types in the dataset. To obtain a mechanistic explanation for these interactions, and to identify the interacting genes between the two cell types we focused on

ligand-target interactions between the two cells types. Specifically, for a predicted directed edge between cell types A and B we learned a Ligand-Target Regression (LTR) model to determine if there is an underlying cell type – cell type interaction between A and B. Our assumption is that if these two cell types indeed interact, then the expression of some of the ligands in cell A type should be able to explain some of the expression changes observed in cell type B. Similar approaches have been used by others to explore cell-type interactions in non case control studies (7).To identify a set of ligands in A predicted to activate or repress target genes in B we optimized the following regression model:

$$\min_{\alpha} \sum_{t}^{T} \left(\sum_{l}^{L} I_{(t,l)} L_{(l)} \alpha_{(l)} - T_{(t)} \right)^2 + \lambda \|\alpha\| \quad (4)$$

Where I represents an input (known) ligand-target interaction matrix (6), L is an input vector of log values for the expression of ligands in cell type A, α represents the (unobserved) ligand activation vector, T represents the expression levels for target genes in cell type B and λ is a regularization parameter. Here we used a $L1$ regularization which usually leads to the selection of relatively few non zero values (corresponding to relatively few activated ligands in cell type A).

Using the inputs to sett $A_{(t,l)} = I_{(t,l)} L_{(l)}$, transforms the optimization problem to

$$\min_{\alpha} \sum_{t}^{T} \left(\sum_{l}^{L} A_{(t,l)} \alpha_{(l)} - T_{(t)} \right)^2 + \lambda \|\alpha\| \quad (5)$$

Which is a standard least absolute shrinkage and selection operator (LASSO) model. To learn parameters for the model we used the "LASSOCV" function from the Python library "scikit-learn", which implements the LASSO cross validation. Note that the model in equation 5 using the same ligand activity parameters for all genes (there is only 1 ligand activity parameter in the model for each ligand across all target genes). Thus, we can use this model in a cross validation setting to predict the expression levels of held out targets in cell type B. For these, we know the ligand-target

interaction from Matrix I and the ligand expression from L allowing us to evaluate the ability of the model to generalize to unseen targets. We also use the model to test if we obtain better prediction accuracy for significant pairs identified in the BNs.

Training and Test for Ligand-Target Regression (LTR) Model

We used a five-fold cross validation strategy to train and test the LTR model: We split the training part of each validation set into two sets to select the hyperparameter λ (our penalty term) and then retrain using all training data for this set and the selected λ to obtain the model used for the fold test data. Evaluation of predicted values is based on the average Pearson correlation between the predicted and actual expression changes for each fold. Following testing we use the average product between the log fold change and coefficient value α in the five-fold training models to rank the list of active ligands.

Joint plots of Bayesian Network and LTR model scores for cell type pairs

To jointly plot the Bayesian network bootstrap score and the Pearson correlation regression score for each cell type pair, we first converted the edge count to log value. For the Pearson correlation we used the average correlation for the five-fold results. For both IPF lung data and lung aging data, cell pairs with edge count smaller than 20 are removed (See **Tab. S12** for details). Note that for some of the pairs we tried to model using LASSO the learning terminated with coefficients of 0 for all ligands (this happened for all runs of the random interaction matrix as we mention in Results and to a few of the CV runs of the cell-cell and intra-cell models). In such cases these models were removed from the correlation analysis.

Comparison to CellPhoneDB, iTALK and NicheNet

All the three methods are based on ligand related gene expression analysis. For CellPhoneDB, the result contains all possible cell type pairs with calculated ligand-receptor scores. For each cell type pair, we use the sum of all its ligand-receptor scores as its pair score. iTALK detects significant ligand- receptor pairs, and provides the mean expression level of them. We then sum the product of ligand and receptor expression as the final score for each cell type pair. NicheNet can select top functional ligand genes with prediction scores for a given cell type pair. We next sum these prediction scores for all selected ligands to rank cell type pairs.

Data Availability

All scripts, and instruction required to run CINS pipeline in Python and R can be found in our support website, https://github.com/xiaoyeye/CINS. Generated sequencing data is available at GEO accession number GSE165638. All other public data can be found following the pipelines in Methods.

Funding

Work partially funded by the National Institutes of Health (NIH) (https://www.nih.gov) [grants 1R01GM122096 and OT2OD026682 to Z.B.J. and 1R01HL127349 to N.K.].

Acknowledgements

None

Author contributions

Y.Y., M.R. and Z.B.-J. designed research; Y.Y., C.C.J., T.S.A., J.S., K.S., N.X., M.R., J.L, N.K., and Z.B.-J. performed research; C.C.J., T.S.A., J.S., K.S., N.X., M.R., N.K. generated data; Y.Y., J.L analyzed data; and Y.Y., C.C.J., and Z.B.-J. wrote the paper. All authors read and approved the paper.

Competing interests

None

1. Deng, Y., Bao, F., Dai, Q., Wu, L.F. and Altschuler, S.J. (2019) Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat Methods*, **16**, 311-314.
2. Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z. and Bar-Joseph, Z. (2018) A web server for comparative analysis of single-cell RNA-seq data. *Nat Commun*, **9**, 4768.
3. Kumar, M.P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D.C., Lauffenburger, D.A. and Raue, A. (2018) Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep*, **25**, 1458-1468 e1454.
4. Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., Huang, H., Yuan, G.C. and Guo, G. (2018) Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol*, **19**, 47.
5. Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M. and Yanai, I. (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol*, **38**, 333-342.
6. Browaeys, R., Saelens, W. and Saeys, Y. (2019) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 1-4.
7. Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-Tormo, R. (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc*, **15**, 1484-1506.
8. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G., Raby, B.A., DeIuliis, G., Januszyk, M. *et al.* (2020) Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv*, **6**, eaba1983.
9. Angelidis, I., Simon, L.M., Fernandez, I.E., Strunz, M., Mayr, C.H., Greiffo, F.R., Tsitsiridis, G., Ansari, M., Graf, E. and Strom, T.-M. (2019) An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature communications*, **10**, 1-17.
10. Hines, P.J. (2019) Stabilizing cell-type ratios. *Science*, **366**, 1210-1210.
11. Willett, R.T., Bayin, N.S., Lee, A.S., Krishnamurthy, A., Wojcinski, A., Lao, Z., Stephen, D., Rosello-Diez, A., Dauber-Decker, K.L. and Orvis, G.D. (2019) Cerebellar nuclei excitatory neurons regulate developmental scaling of presynaptic Purkinje cell number and organ growth. *Elife*, **8**, e50617.
12. Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I. and Linnarsson, S. (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature methods*, **15**, 932.
13. Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S., Li, C. and Amamoto, R. (2014) Highly multiplexed subcellular RNA sequencing in situ. *Science*, **343**, 1360-1363.
14. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A. and Dulac, C. (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, **362**, eaau5324.
15. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O. and Huss, M. (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78-82.
16. Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C. and Yuan, G.-C. (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, **568**, 235.
17. Xia, C., Fan, J., Emanuel, G., Hao, J. and Zhuang, X. (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, **116**, 19490-19499.

18. Yuan, Y. and Bar-Joseph, Z. (2020) GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biol*, **21**, 300.
19. Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G. and Bar-Joseph, Z. (2019) Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, **7**, 54.
20. Imoto, S., Kim, S.Y., Shimodaira, H., Aburatani, S., Tashiro, K., Kuhara, S. and Miyano, S. (2002) Bootstrap analysis of gene networks based on Bayesian networks and nonparametric regression. *Genome Informatics*, **13**, 369-370.
21. Kia'i, N. and Bajaj, T. (2020), *StatPearls*, Treasure Island (FL).
22. Qiu, H., He, Y., Ouyang, F., Jiang, P., Guo, S. and Guo, Y. (2019) The Role of Regulatory T Cells in Pulmonary Arterial Hypertension. *J Am Heart Assoc*, **8**, e014201.
23. Noubade, R., Majri-Morrison, S. and Tarbell, K.V. (2019) Beyond cDC1: Emerging Roles of DC Crosstalk in Cancer Immunity. *Front Immunol*, **10**, 1014.
24. Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., Wang, M., Ajani, J., Futreal, A. and Wang, L. (2019) iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *bioRxiv*, 507871.
25. Dickey, B.F. and Whitsett, J.A. (2017) Understanding Interstitial Lung Disease: It's in the Mucus. *Am J Respir Cell Mol Biol*, **57**, 12-14.
26. Platonova, N., Miquel, G., Regenfuss, B., Taouji, S., Cursiefen, C., Chevet, E. and Bikfalvi, A. (2013) Evidence for the interaction of fibroblast growth factor-2 with the lymphatic endothelial cell marker LYVE-1. *Blood, The Journal of the American Society of Hematology*, **121**, 1229-1237.
27. Avraham, T., Daluvoy, S., Zampell, J., Yan, A., Haviv, Y.S., Rockson, S.G. and Mehrara, B.J. (2010) Blockade of transforming growth factor-beta1 accelerates lymphatic regeneration during wound repair. *Am J Pathol*, **177**, 3202-3214.
28. Doucet, C., Brouty-Boye, D., Pottin-Clemenceau, C., Canonica, G.W., Jasmin, C. and Azzarone, B. (1998) Interleukin (IL) 4 and IL-13 act on human lung fibroblasts. Implication in asthma. *J Clin Invest*, **101**, 2129-2139.
29. Coker, R.K., Laurent, G.J., Shahzeidi, S., Lympany, P.A., du Bois, R.M., Jeffery, P.K. and McAnulty, R.J. (1997) Transforming growth factors-beta 1, -beta 2, and -beta 3 stimulate fibroblast procollagen production in vitro but are differentially expressed during bleomycin-induced lung fibrosis. *Am J Pathol*, **150**, 981-991.
30. Kara Rogers Senior Editor, B.S. (2010) *The Respiratory System*. Britannica Educational Pub.
31. Bergamaschi, E., Canu, I.G., Prina-Mello, A. and Magrini, A. (2017), *Adverse effects of engineered nanomaterials*. Elsevier, pp. 125-158.
32. Acton, Q.A. (2013) *Asthma: New Insights for the Healthcare Professional: 2013 Edition*. ScholarlyEditions.
33. Sabbione, F., Gabelloni, M.L., Ernst, G., Gori, M.S., Salamone, G., Oleastro, M., Trevani, A., Geffner, J. and Jancic, C.C. (2014) Neutrophils suppress gammadelta T-cell function. *Eur J Immunol*, **44**, 819-830.
34. Rodriguez-Pinto, D., Saravia, N.G. and McMahon-Pratt, D. (2014) CD4 T cell activation by B cells in human Leishmania (Viannia) infection. *BMC Infect Dis*, **14**, 108.
35. Salam, N., Rane, S., Das, R., Faulkner, M., Gund, R., Kandpal, U., Lewis, V., Mattoo, H., Prabhu, S., Ranganathan, V. *et al.* (2013) T cell ageing: effects of age on development, survival & function. *Indian J Med Res*, **138**, 595-608.
36. Grivennikov, S.I., Tumanov, A.V., Liepinsh, D.J., Kruglov, A.A., Marakusha, B.I., Shakhov, A.N., Murakami, T., Drutskaya, L.N., Forster, I., Clausen, B.E. *et al.* (2005) Distinct and nonredundant in vivo functions of TNF produced by t cells and macrophages/neutrophils: protective and deleterious effects. *Immunity*, **22**, 93-104.

37. Lahn, M., Kalataradi, H., Mittelstadt, P., Pflum, E., Vollmer, M., Cady, C., Mukasa, A., Vella, A.T., Ikle, D., Harbeck, R. *et al.* (1998) Early preferential stimulation of gamma delta T cells by TNF-alpha. *J Immunol*, **160**, 5221-5230.
38. (2020).
39. Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thurmann, L., Kurth, F., Volker, M.T. *et al.* (2020) COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat Biotechnol*, **38**, 970-979.
40. Morgado, F.N., da Silva, A.V.A. and Porrozzi, R. (2020) Infectious Diseases and the Lymphoid Extracellular Matrix Remodeling: A Focus on Conduit System. *Cells*, **9**.
41. Kumagai, Y., Takeuchi, O., Kato, H., Kumar, H., Matsui, K., Morii, E., Aozasa, K., Kawai, T. and Akira, S. (2007) Alveolar macrophages are the primary interferon-alpha producer in pulmonary infection with RNA viruses. *Immunity*, **27**, 240-252.
42. Chaudhuri, V. and Karasek, M.A. (2006) Mechanisms of microvascular wound repair II. Injury induces transformation of endothelial cells into myofibroblasts and the synthesis of matrix proteins. *In Vitro Cell Dev Biol Anim*, **42**, 314-319.
43. Gochhait, D., Dey, P. and Verma, N. (2016) Cytology of plasma cell rich effusion in cases of plasma cell neoplasm. *J Cytol*, **33**, 150-153.
44. Smyth, L.C.D., Rustenhoven, J., Scotter, E.L., Schweder, P., Faull, R.L.M., Park, T.I.H. and Dragunow, M. (2018) Markers for human brain pericytes and smooth muscle cells. *J Chem Neuroanat*, **92**, 48-60.
45. Sweeney, M. and Foldes, G. (2018) It Takes Two: Endothelial-Perivascular Cell Cross-Talk in Vascular Development and Disease. *Front Cardiovasc Med*, **5**, 154.

Tab. 1 Top differential cell type interactions identified by CINS for the IPF dataset. The IPF-Control column lists the difference in the number of times the edge between the two cells was identified in 100 bootstrap runs for each of the two datasets. Negative values indicate that it was identified more for the Control whereas positive numbers mean that the interaction is more prevalent in IPF. For all listed edges the interaction was only identified in for one of the two datasets (score of 100 or -100).

cell_type1	cell_type2	IPF-Control	Reference
Macrophage	Ciliated	-100	There is strong interaction between ciliated cell and Macrophage in COVID-19 critical cases (39)
Fibroblast	Lymphatic	-100	Fibroblast produce extracellular matrix which is critical to lymph node microenvironment (40)
cDC2	DC_Mature	100	
cDC2	cDC1	-100	cDC2 and cDC1 are cross-talking with each other (23)
Macrophage	cDC1	100	
Mesothelial	Aberrant_Basaloid	100	
Macrophage_Alveolar	pDC	-100	Macrophage_Alveolar (AM) and pDC are involved in antiviral immune, and pDC will be activated if the AM defense line is broken (41)
Myofibroblast	VE_Venous	-100	Injury lets endothelial cells transform to myofibroblast (42)
Ciliated	ncMonocyte	-100	Ciliated cells may contribute to monocyte inflow in COVID-19 (39)
Multiplet	VE_Capillary_B	100	
B_Plasma	Mesothelial	-100	Excess plasma cells are found with mesothelial cells on effusion cytology smear (43)
VE_Capillary_B	SMC	-100	
Pericyte	SMC	100	Brain pericytes and vascular SMC comprise mural cells which is important to support blood vessels (44)
ncMonocyte	Multiplet	100	

ncMonocyte	DC_Mature	-100	
T_Regulatory	Fibroblast	100	Treg cell regulates fibroblast in lung (22)
VE_Arterial	VE_Venous	100	
T	T_Regulatory	100	
VE_Peribronchial	Pericyte	100	One pericyte can communicate with more than one endothelial cells (45)
T_Regulatory	DC_Langerhans	-100	

Figure Legends

Figure 1. Overview of CINS. (A) Cell type annotation is used to extract cell type fractions in each sample. Next cell type fraction is discretized by learning Gaussian Mixture Model (GMM) for this type, respectively. (B) A Bayesian network (BN) is learned using the discretized cell abundance information. Bootstrapping is performed to identify high scoring differential interactions between cell types. (C) For pairs identified in the directed bootstrap BN analysis, a ligand-target regression (LTR) model is learned. In this model we use expression of ligands in the cell type with the outgoing edge to predict the expression of targets genes in the cell type with incoming edge. (D) Finally, LTR is used to select key ligands that underlie the cell-cell interactions identified in the BN. cell interaction.

Figure 2. Bayesian Networks (BN) learned for lung cell types in healthy and IPF individual. (A) BN for controls (healthy individuals). (B) BN for IPF patients. Nodes represent specific cell

types and are colored accordingly, edges represent directed interactions between the cell types. Edge width corresponds to its bootstrap score.

Figure 3. Interactions learned by the BN are more significant than interactions between cells of the same type. Comparison between the ability of the LTR model to predict target expression when learning the model using cell pairs identified by the BN (A) and the same cell type (B). The x axis represents the bootstrapped edge count (score) of the interaction in the BN for a cell type pair, and the y axis represents the LTR model performance (higher is better) for the same cell pair.

Figure 4. Aging Bayesian Networks. (A) BN for young mice. B) BN for adult mice. Nodes and edges notations and colorings are similar to those used in **Fig. 2**.

Figure 5. LTR comparison for the aging data. Comparison between the ability of the LTR model to predict target expression when learning the model using cell pairs identified by the BN (A) and the same cell type (B).

Figure 6. Permutation analysis highlights the agreement between the two aging networks. (A) Leftmost – learning using the Angelidis (15 samples) dataset. (B) Top – Learning combined networks using both Angelidis and real new data. Bottom – Learning combined networks using both Angelidis and permutation of cell type fractions in the new data. (C) Overlap in bootstrapped

edges between the original and combined model when using the real data (red dashed line) and the permutation data (blue distribution).