


Gene expression

# Dysregulated ligand–receptor interactions from single-cell transcriptomics

Qi Liu <sup>1,2,†</sup>, Chih-Yuan Hsu<sup>1,2,†</sup>, Jia Li<sup>1,2</sup> and Yu Shyr<sup>1,2,\*</sup>

<sup>1</sup>Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37232, USA and <sup>2</sup>Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Can Alkan

Received on June 16, 2021; revised on March 29, 2022; editorial decision on April 18, 2022; accepted on April 21, 2022

## Abstract

**Motivation:** Intracellular communication is crucial to many biological processes, such as differentiation, development, homeostasis and inflammation. Single-cell transcriptomics provides an unprecedented opportunity for studying cell-cell communications mediated by ligand–receptor interactions. Although computational methods have been developed to infer cell type-specific ligand–receptor interactions from one single-cell transcriptomics profile, there is lack of approaches considering ligand and receptor simultaneously to identifying dysregulated interactions across conditions from multiple single-cell profiles.

**Results:** We developed scLR, a statistical method for examining dysregulated ligand–receptor interactions between two conditions. scLR models the distribution of the product of ligands and receptors expressions and accounts for inter-sample variances and small sample sizes. scLR achieved high sensitivity and specificity in simulation studies. scLR revealed important cytokine signaling between macrophages and proliferating T cells during severe acute COVID-19 infection, and activated TGF- $\beta$  signaling from alveolar type II cells in the pathogenesis of pulmonary fibrosis.

**Availability and implementation:** scLR is freely available at <https://github.com/cyhsuTN/scLR>.

**Contact:** [yu.shyr@vumc.org](mailto:yu.shyr@vumc.org)

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Cells constantly communicate with each other to orchestrate their behaviors, ensuring normal functions of tissues, organs and bodies. One important mode of intercellular communication is ligand–receptor interactions, where ligands from a ‘sender’ cell bind to receptors in a ‘receiver’ cell that triggers response inside the cell. The recent advance of single-cell RNA sequencing technology provides a powerful way to study ligand–receptor interactions at an unprecedented scale and depth (Almet *et al.*, 2021; Armingol *et al.*, 2021). Deciphering ligand–receptor interactions from single-cell transcriptomics has become routine for understanding the biological systems (Martin *et al.*, 2019).

Computational methods have been developed rapidly to infer cell-cell communications mediated by ligand–receptor interactions from single-cell transcriptomics (Browaeys *et al.*, 2020; Cabello-Aguilar *et al.*, 2020; Efremova *et al.*, 2020; Hou *et al.*, 2020; Hu *et al.*, 2021; Jin *et al.*, 2021; Noel *et al.*, 2021; Solovey and Scialdone, 2020; Zhang *et al.*, 2021). Ligand–receptor interactions are detected and quantified based on the pairwise expression of

ligands and receptors between single cells or cell clusters. Some methods, such as ICELLNET (Noel *et al.*, 2021), determine the interaction strength by the product of ligands and receptors expressions. Other methods, such as CellChat (Jin *et al.*, 2021) and SingleCellSignalR (Cabello-Aguilar *et al.*, 2020), calculate the interaction by a non-linear transformation of the product of ligands and receptors expressions. The significance of cell type-specific interactions is estimated based on a null distribution generated by shuffling cluster labels of all cells. These methods have been successfully applied in a number of single-cell transcriptomics datasets, which uncovered key signaling mechanisms controlling cell fate and state transitions (Bassez *et al.*, 2021; Cheng *et al.*, 2021; Gong *et al.*, 2021; Hildreth *et al.*, 2021; Tian *et al.*, 2021).

Existing methods focus primarily on the inference of ligand–receptor interactions between cells/cell clusters in one condition. The identification of dysregulated communications across conditions, however, is even more important for revealing potential driving signals. iTalk and CellChat predict up- (gained) or down-regulated (loss) interactions based on the differentially expressed ligands and/or receptors, where upregulated/gained interactions are defined

based on upregulated ligands and/or receptors and vice versa. The differential analysis compares expressions from pooled cells between two conditions, which fail to consider inter-sample variances in each condition. Moreover, identification of dysregulated ligand–receptor interactions based on expression changes of ligands or receptors alone would result in false positives and false negatives. For example, a ligand–receptor interaction with an upregulated ligand but a downregulated receptor is probably not a strong candidate.

We developed scLR to identify dysregulated ligand–receptor interactions across conditions. scLR not only models the distribution of the product of ligands and receptors expressions, but also accounts for inter-sample variances, small sample sizes and dropout events. scLR achieved high recall (sensitivity) and specificity in four simulation datasets. scLR revealed important cytokine signaling between macrophages and proliferating T cells in severe COVID-19 infection, and activated TGF- $\beta$  signaling from alveolar type II cells in the pathogenesis of pulmonary fibrosis.

## 2 Materials and methods

### 2.1 Data preprocessing

The inputs of scLR are one raw gene–cell count matrix along with the cell cluster and the sample information (sample ID and condition) from single-cell RNA-seq data. The sample represents a replicate, such as one patient. For example, the sample size is 20 if the single-cell RNA-seq dataset is generated from 10 tumor patients and 10 controls. scLR first sums gene counts in each cell cluster in each sample to estimate gene expression in the cluster and in the sample. Then scLR normalizes the summed data using a median normalization method in DESeq2, which assumes expression profiles have the same median value (Anders and Huber, 2010) (details in Normalization in the Supplementary File). After normalization, scLR transforms the data by  $\log_2(x+1)$  to obtain expression abundances for the ligand  $y_L^{c,s}$  and the receptor  $y_R^{c,s}$  in the cluster  $c$  of sample  $s$ , where  $c \in \{c_1, c_2, \dots, c_m\}$  and  $s \in \{s_1, s_2, \dots, s_n\}$ .  $m$  is the number of clusters and  $n$  is the number of samples.

### 2.2 The distribution of the product of ligand and receptor expressions

We assume  $y_L^{c,s}$  and  $y_R^{c,s}$  follow a bivariate normal distribution with correlation  $\rho$ , where  $y_L^{c,s}$  and  $y_R^{c,s}$  denote the expression of ligand and receptor in the cluster  $c_i$  and  $c_j$  of sample  $s$ , respectively. To simplify the equation, we use  $y_L$  and  $y_R$  to represent  $y_L^{c_i,s}$  and  $y_R^{c_j,s}$  (Equation 1).

$$\begin{pmatrix} y_L \\ y_R \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_L \\ \mu_R \end{pmatrix}, \begin{pmatrix} \sigma_L^2 & \rho\sigma_L\sigma_R \\ \rho\sigma_L\sigma_R & \sigma_R^2 \end{pmatrix}\right) \quad (1)$$

$y_L y_R$ , the product of  $y_L$  and  $y_R$ , can be expressed as (Equation 2).

$$y_L y_R = \frac{\sigma_L \sigma_R}{4} \left[ 2(1+\rho)\chi_{1,a}^2 - 2(1-\rho)\chi_{1,b}^2 \right] \quad (2)$$

where  $\chi_{1,a}^2 = \left(\frac{y_L + y_R}{\sigma_L + \sigma_R}\right)^2$  and  $\chi_{1,b}^2 = \left(\frac{y_L - y_R}{\sigma_L - \sigma_R}\right)^2$ .  $\chi_{1,a}^2$  and  $\chi_{1,b}^2$  are independent, which follow non-central chi-square distributions with the degree of freedom of 1 and the non-central parameters of  $a =$

$\left(\frac{\mu_L + \mu_R}{\sigma_L + \sigma_R}\right)^2$  and  $b = \left(\frac{\mu_L - \mu_R}{\sigma_L - \sigma_R}\right)^2$ , respectively. As shown in (2), the distribution of  $y_L y_R$  is equal to that of the difference of two independent non-central chi-square random variables (regardless of constants), which depends on  $\sigma_L \sigma_R$ ,  $a$ ,  $b$ , and  $\rho$ . The expectation  $E(y_L y_R) = \mu_L \mu_R + \rho \sigma_L \sigma_R$ .

### 2.3 Differential analysis of ligand–receptor interactions

scLR is aimed to identify cell-type specific dysregulated interactions between two conditions. It first calculates the interaction strength of each ligand–receptor pair between two cell clusters (the ‘sender’

and ‘receiver’) and then assesses its alteration between two conditions statistically. The interaction strength of ligand–receptor between clusters  $c_i$  and  $c_j$  in sample  $s$  is measured by the product of the ligand and receptor in the corresponding cluster  $c_i$  and  $c_j$  of sample  $s$ :  $y_L^{c_i,s} y_R^{c_j,s}$ . Assume there are  $n_A$  samples from condition A and  $n_B$  samples from condition B, where  $n = n_A + n_B$ . The mean interaction strength of ligand and receptor between clusters  $c_i$  and  $c_j$  across all the samples under condition A can be denoted as  $\overline{y_L^{c_i,A} y_R^{c_j,A}}$ , where  $\overline{y_L^{c_i,A} y_R^{c_j,A}} = \frac{1}{n_A} \sum_{s_k \in A} y_L^{c_i,s_k} y_R^{c_j,s_k}$ . Similarly,  $\overline{y_L^{c_i,B} y_R^{c_j,B}} = \frac{1}{n_B} \sum_{s_k \in B} y_L^{c_i,s_k} y_R^{c_j,s_k}$ . We compare the mean interaction strengths of ligand–receptor interactions under two different conditions.

The null hypothesis is that the mean of ligand–receptor interaction strengths are the same under two conditions. The alternative hypothesis is that they are not the same.

$$\begin{aligned} H_0 &: |E(y_L^{c_i,A} y_R^{c_j,A}) - E(y_L^{c_i,B} y_R^{c_j,B})| = 0 \\ H_1 &: |E(y_L^{c_i,A} y_R^{c_j,A}) - E(y_L^{c_i,B} y_R^{c_j,B})| \neq 0 \end{aligned}$$

If the difference  $|d| = \left| \overline{y_L^{c_i,A} y_R^{c_j,A}} - \overline{y_L^{c_i,B} y_R^{c_j,B}} \right| > c_{1-\alpha}$  for some specified  $c_{1-\alpha}$ , we reject the null hypothesis and conclude that ligand and receptor interactions are significantly altered between two conditions.

We use Monte Carlo simulation to estimate the null distribution of  $|d|$ . We first estimate the means and standard deviations of the ligand and receptor expressions in clusters  $c_i$  and  $c_j$  under two conditions A and B, denoted by  $\hat{\mu}_{L,A}$ ,  $\hat{\mu}_{L,B}$ ,  $\hat{\mu}_{R,A}$ ,  $\hat{\mu}_{R,B}$ ,  $\hat{\sigma}_{L,A}$ ,  $\hat{\sigma}_{L,B}$  and  $\hat{\sigma}_{R,A}$ ,  $\hat{\sigma}_{R,B}$ . To assess inter-sample variances accurately when the sample size is small, we use a Bayesian method to shrink the estimated variances toward a pooled estimate (Smyth, 2004), which is performed by eBayes in the R package limma. Moreover, we assume  $\rho = 0$  based on the observation of a very low percentage of significant non-zero correlations in real datasets, e.g. only 0.05% of ligand–receptor pairs in the COVID-19 data, 0.6% pairs were found in the IPF data and 0.03% were detected in AVP data at adjusted  $P$ -value  $< 0.05$ . The scLR package also provides a function to estimate  $\rho$  (default = 0) in case there are a number of significant non-zero correlations. When the null hypothesis is true,  $|E(y_L^{c_i,A} y_R^{c_j,A}) - E(y_L^{c_i,B} y_R^{c_j,B})| = 0$ , which can be simplified as  $|E(y_L^{c_i,A} y_R^{c_j,A}) - E(y_L^{c_i,B} y_R^{c_j,B})| = 0$  between clusters  $c_i$  and  $c_j$ , we can assume  $\sigma_{L,A} \sigma_{R,A} = \sigma_{L,B} \sigma_{R,B}$ ,  $a_A = a_B$  and  $b_A = b_B$  based on Equation (2), where  $a_A$ ,  $a_B$ ,  $b_A$  and  $b_B$  are the non-central parameters of non-central chi-square distributions in conditions A and B. We let  $\sigma_L \hat{\sigma}_R = w \hat{\sigma}_{L,A} \hat{\sigma}_{R,B} + (1-w) \hat{\sigma}_{L,B} \hat{\sigma}_{R,B}$ ,  $\hat{a} = w \hat{a}_A + (1-w) \hat{a}_B$  and  $\hat{b} = w \hat{b}_A + (1-w) \hat{b}_B$ , where  $w = n_A / (n_A + n_B)$ . Based on  $\sigma_L \hat{\sigma}_R$ ,  $\hat{a}$  and  $\hat{b}$ , we simulate the distribution of  $y_L y_R$  based on the Equation (2). We randomly choose  $n_A$  samples to estimate  $\overline{y_L^{c_i,A^*} y_R^{c_j,A^*}}$  and  $n_B$  samples to derive  $\overline{y_L^{c_i,B^*} y_R^{c_j,B^*}}$ . The difference between two conditions  $d^* = \overline{y_L^{c_i,A^*} y_R^{c_j,A^*}} - \overline{y_L^{c_i,B^*} y_R^{c_j,B^*}}$ . When we repeat the simulation  $K$  times, we have  $\{d_1^*, \dots, d_K^*\}$ . The  $c_{1-\alpha}$  required is the 100(1- $\alpha$ )% percentile of  $\{|d_1^*|, \dots, |d_K^*|\}$  and the  $P$ -value is given by the proportion of  $\{d_1^*, \dots, d_K^* : |d_i^*| > |d|\}$ .

A decent number of simulations is needed to obtain an accurate estimation of the significance of dysregulated ligand–receptor interactions. It is very time consuming, however, if a large number of simulations are conducted for every ligand–receptor pair between any two cell clusters since there are thousands of ligand–receptor pairs and tens of cell clusters. To address this problem, we estimate the density of  $\{d_1^*, \dots, d_K^*\}$  using a kernel density approach with the Gaussian kernel and a large data-driven bandwidth  $h$  in each ligand–receptor pair comparison. The bandwidth  $h$  controls smoothness of the estimated density. A small  $h$  gives very rough estimates while a large  $h$  gives smoother estimates but weakens local features, for example, smoothing tails but weakening the mode(s) of the density. The main purpose here is to obtain an accurate tail probability rather than an overall density estimate. We use a data-driven

approach to estimate  $b$  via the R function *density* (parameters:  $\text{bw} = \text{'SJ'}$  and  $\text{adjust} = 3$ ).

Based on the estimated density  $\hat{f}(x) = \frac{1}{K} \sum_{i=1}^K \frac{1}{\sqrt{2\pi}b} \exp\left\{-\frac{1}{2}\left(\frac{x-d_i}{b}\right)^2\right\}$ ,  $-\infty < x < \infty$ , the  $P$ -value is estimated by two times the tail probability beyond the  $d$ .

## 2.4 Pseudo-counts for reducing false positives

The prevalence of dropout events is one of the greatest challenges in single-cell RNA sequencing data analysis, which introduces excessive number of zeros and unwanted technical variability. There is a close relationship between dropout rate and expression level, where genes with high dropout rate indicate low expression (Qiu, 2020). The excessive number of zeros make ligand–receptor interaction analysis even more difficult. The product would be zero if the ligand or the receptor expression in one condition are undetected, which leads to many false positives especially when sample sizes are small. To reduce the number of false positives caused by dropout events, we add pseudo-count of 1 to the zero values. Comparing scLR with and without pseudo-counts using simulation data, we found that scLR with pseudo-counts achieved higher precision and specificity (i.e. lower false positive rates) (Supplementary Fig. S1). The scLR package also provides a parameter (`zero.impute, default=TRUE`) to turn off pseudo-counts.

## 2.5 Single-cell RNA-seq datasets

We used three single-cell RNA-seq datasets to evaluate the performance of scLR, including data generated from patients with and without COVID-19 infection (Grant et al., 2021), pulmonary fibrosis lungs (PF) and non-fibrotic controls (Habermann et al., 2020), and patients with Anterior vaginal prolapse (AVP) and controls (Li et al., 2021). Three datasets were all generated by 10X platform. The COVID-19 data were downloaded from GSE155249, including five patients with severe COVID-19 infection and two COVID-19 negatives. There were 33 000 cells with 18 337 genes, clustered into six cell types, macrophages, dendritic cells (DC), T cells, proliferating T cells, ciliated and alveolar cells. The PF data were downloaded from GSE135893, consisting of 11 PF patients and nine controls. There were 70 512 cells with 33 694 genes, classified into 12 cell types. The AVP data were downloaded from GSE151202, involving 16 AVP patients and five controls. There were 53 133 cells with 20 328 genes, categorized into seven cell types. Multivariate normality test showed that only 2.1% of ligand–receptor pairs were significant different from bivariate normal distributions in the COVID-19 data, 12.1% pairs in IPF data and 1.6% in AVP data at an adjusted  $P$ -value  $< 0.05$ . These results suggested that a small percentage of ligand–receptor pairs might be modeled poorly by bivariate normal distributions. For those pairs violating normality assumption, we found that rectified normal distributions and gamma distributions fitted the data well, especially rectified normal distributions.

## 3 Results

### 3.1 Simulation

#### 3.1.1 Simulation settings

We designed four scenarios to evaluate the performance of scLR (Table 1). Each scenario has three samples in each of two

**Table 1.** Simulation settings in four scenarios

|            | Ligand    | Receptor  |
|------------|-----------|-----------|
| Scenario 1 | (↑, ↑, ↑) | (↑, ↑, ↑) |
| Scenario 2 | (↑, ↑, ↑) | (−, −, −) |
| Scenario 3 | (−, −, −) | (↑, ↑, ↑) |
| Scenario 4 | (↑, ↑, ↑) | (↓, ↓, ↓) |

Note: Three samples in each condition.

conditions. Both ligands and receptors are upregulated in the first setting, whereas only ligands or receptors are upregulated in the second scenario. The third setting is challenging, where ligands and receptors are overexpressed complementarily, e.g. highly upregulated ligands but slightly or non-upregulated receptors in one sample and vice versa in the other sample, leading to increased products of ligand and receptor expressions. Upregulated interactions in this scenario are not driven by ligands or receptors alone, but the complementary increase of ligands and receptors. We simulated complementary increase by upregulation of ligands but slight overexpression of receptors in the samples 1 and 3, and upregulation of receptors but slight overexpression of ligands in the sample 2. The fourth scenario is a special case, where ligands are upregulated but receptors are downregulated, resulting in unchanged ligand–receptor interactions. There are 2000 ligand–receptor pairs in each setting. There are 500 significant and 1500 non-significant pairs in the first three scenarios. There are no true positives but 2000 non-significant pairs in the fourth scenario. Every sample in the simulation studies consists of 16 672 genes and 11 cell clusters. The expression abundances were generated from normal distributions, where the parameters of means and standard deviations for each gene in each cell cluster were derived from real single-cell RNA-seq datasets. The normalized counts were increased/decreased to simulate the upregulation/downregulation on the expression abundances (see Simulation Settings in the Supplementary File). Ten simulated datasets were generated for each scenario.

#### 3.1.2 Performance on simulation studies

We compared scLR with Welch's  $t$ -test and differential gene expression analysis by limma under four simulation settings. Welch's  $t$ -test was used to compare the mean difference of the product of ligands and receptors expressions between two conditions based on the assumption that the product follows a normal distribution. Limma was performed to identify differential expression in ligands or receptors, where ligand–receptor pairs with either ligands or receptors differentially expressed between two conditions were considered as dysregulated ligand–receptor interactions.

We evaluated the performances based on the following metrics: (i) the F1 score curve at different thresholds; (ii) the recall (sensitivity) and specificity curve at different thresholds; and (iii) the number of true positives (TP), precision (TP/TP+FP), recall (TP/TP+FN), specificity (TN/TN+FP) and F1 score at the threshold of adjusted  $P$ -value  $< 0.01$ .

In the first scenario where dysregulated ligand–receptor interactions were caused by both upregulation in the ligand and receptor expressions, scLR showed the best performance. scLR achieved much higher F1 score and recall than the other two methods at different thresholds (Fig. 1A). The three methods obtained similar specificity (Fig. 1A). At the threshold of adjust  $P$ -value  $< 0.01$ , scLR recovered ~430 out of 500 significant ligand–receptor pairs (84% recall), while  $t$ -test only identified ~100 pairs (20% recall) and limma detected ~330 pairs (66% recall) (Supplementary Fig. S2). When there were an increasing percentage of pairs (ranging from 0% to 90%) violating the normality assumption (either from Gamma or rectified normal distributions), the specificity of scLR was controlled well but the recall decreased (Supplementary Fig. S3). scLR showed much better performance in terms of recall and F1 scores than limma and  $t$ -test when there were 50% of non-normal pairs (Gamma or rectified normal distributions), indicating the power of scLR decreased less than limma and  $t$ -test as the percentage of non-normal pairs increased (Supplementary Fig. S4).

In the second scenario where dysregulated ligand–receptor interactions were driven by either the ligand or the receptor alterations alone, scLR had slightly inferior performance than limma (Supplementary Fig. S5). The F1 and recall curve at different thresholds showed that scLR obtained better performance at rigorous cutoffs ( $0.001 < \text{adjusted } P\text{-value} < 0.01$ ), while limma had better performance at relaxed cutoffs ( $0.01 \leq \text{adjusted } P\text{-value} < 0.1$ ) (Fig. 1B).

The third scenario is challenging, where dysregulated ligand–receptor interactions are driven by complementary ligand and receptor

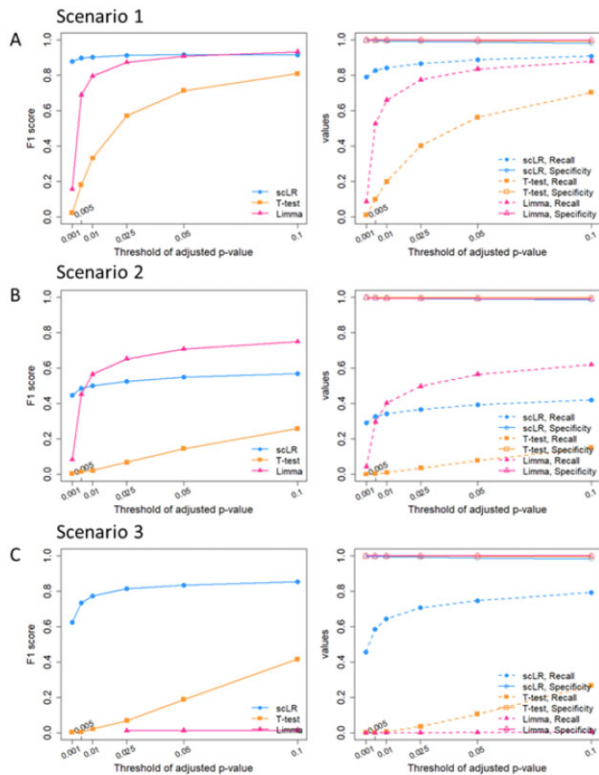


Fig. 1. Performance evaluation in simulation studies. Performance in the first scenario (A), the second scenario (B) and the third scenario (C).

alterations. scLR achieved the best performance, which had much higher F1 score and recall than *t*-test and limma at different thresholds (Fig. 1C). limma, relying on differential expression of either ligands or receptors alone, showed poor performance. scLR identified ~330 interacting pairs with 64% recall, whereas *t*-test and limma both failed to detect any ligand–receptor pairs at the threshold of adjusted *P*-value < 0.01 (Fig. 1C and Supplementary Fig. S6).

The fourth scenario simulated a special case where none of dysregulations exist due to opposite changes in ligands and receptors expression. In this case, scLR and *t*-test performed better than limma, which obtained much lower false positives rates (Supplementary Fig. S7).

The simulation studies demonstrated that scLR overall achieved higher performance than limma and *t*-test. Similar results were also obtained in the same four scenarios with five samples in each condition (Supplementary Figs S8–S11). Since dysregulated ligand–receptor interactions involve both ligands and receptors, methods considering ligands or receptors alone like limma are not effective, such as in scenarios 1, 3 and 4. *T*-test considers ligand–receptor interactions in terms of their products, but it is conservative due to both large estimates for the standard deviations of the products and the inappropriate normal distribution assumption for non-normal products (see Equation 2). scLR directly models the distribution of the product of ligands and receptors expressions, therefore, it achieved much better performance than *t*-test.

### 3.2 Identification of dysregulated ligand–receptor interactions in severe COVID-19

We applied scLR on single-cell RNA-seq data on BAL fluid collected from patients with severe SARS-COV-2 pneumonia and two control patients, one with bacterial pneumonia and one non-pneumonia control (Grant *et al.*, 2021). The integrated analysis of five patients with severe SARS-COV-2 pneumonia and two patients without SARS-COV-2 pneumonia resolved multiple clusters corresponding

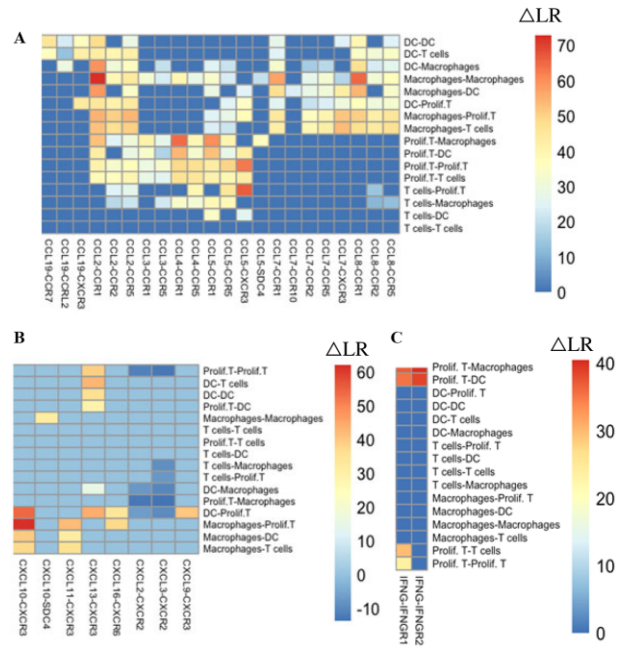


Fig. 2. Dysregulated ligand–receptor interactions in SARS-COV-2 infection compared to control. Dysregulated CC chemokine interactions (A), CXC chemokine interactions (B) and IFNG interactions (C).  $\Delta$ LR is the difference of expression products of ligands and receptors between two conditions

to macrophages, dendritic cells, T cells, proliferating T cells, ciliated and alveolar cells.

Using the curated ligand–receptor database (LRdb) compiled by SingleCellSignalR (Cabello-Aguilar *et al.*, 2020), scLR identified 1313 dysregulated ligand–receptor interactions in SARS-COV-2 infection compared to non-SARS-COV-2 infection (adjusted *P*-value < 0.01) (Supplementary Table S1), of which 938 upregulated and 375 downregulated pairs. Among them, 16% of interactions did not have significantly altered ligands or receptors even at a loose cut off (adjusted *P*-value < 0.1) (Supplementary Fig. S12). Those interactions generally ranked lower than those with significantly altered ligands or receptors (ranking by adjusted *P*-values). Macrophages have the most disrupted autocrine ligand–receptor interactions (128), followed by 88 altered interactions from proliferating T cells to macrophages and 81 from macrophages to proliferating T cells, suggesting active communications between macrophage and proliferating T cells during SARS-COV-2 infection. The most upregulated ligand–receptor pairs involve several cytokines and chemokines that are important for T cell and monocytes recruitment and alveolar macrophage maturation, such as CCL2, CCL3, CCL4, CCL5, CCL7, CCL8, CXCL13, CXCL10, CXCL11 and CXCL16. Macrophages is the major source of dysregulated interactions involving CCL2, CCL7 and CCL8, while proliferating T cell is the source for CCL4, CCL3 and CCL5 (Fig. 2A). CCL2-CCR1 interaction is most upregulated across all cell type pairs, especially between macrophage itself. In the CXC family, CXCL10-CXCR3 is most upregulated in the macrophage-proliferating T cells communication (Fig. 2B). Recent studies reveal that innate immune interferons dysregulation is key to determine SARS-COV-2 pathogenesis. scLR detected the dysregulated IFNG-IFNGR1 and IFNG-IFNGR2 between proliferating T cells and other immune cells, especially between proliferating T and macrophages (Fig. 2C). These findings are consistent with recent studies reporting that macrophages drove the inflammatory response to SARS-COV-2 infection (Speranza *et al.*, 2021) and macrophages and T cells form a positive feedback loop that derives persistent alveolar infection (Grant *et al.*, 2021). In addition, scLR found that it is proliferating T cells that produce IFNG to induce inflammatory cytokine release from macrophages, such as CCL2,

CCL7, CCL8, CXCL10 and CXCL16, which further promote T cell activation and proliferation.

We compared scLR with CellChat and iTALK. Instead of considering ligands and receptors simultaneously, CellChat and iTALK detect dysregulated interactions based on expression changes of either ligands or receptors alone. Additionally, scLR calculates the difference at the sample-level, while CellChat and iTALK pool cells from all samples in one condition together to estimate the difference, which ignore sample-level variances and might lead to biased results. Using the same ligand–receptor database LRdb (Cabello-Aguilar et al., 2020), CellChat identified 1720 dysregulated interactions ( $|\log_2FC| > 0.2$  &  $P$ -value  $< 0.01$ ) (Supplementary Table S2) and iTALK discovered 4385 interactions ( $|\log_2FC| > 0.2$  & adjusted  $P$ -value  $< 0.01$ ) (Supplementary Table S3). We focused on major findings from the original study to further investigate those interactions, which revealed that T cells produce IFNG to induce inflammatory cytokine release and further promote T cell activation (Grant et al., 2021). We listed IFNG-related interactions between immune cells by scLR, CellChat and iTALK in the Supplementary Figure S13, along with differential expression results of ligands and receptors at the patient-level (scLR) and at the cell-level (CellChat and iTALK). scLR found IFNG-related interactions all upregulated after COVID-19 infection, including IFNG-IFNGR1 between proliferating T and all other cell types and IFNG-IFNGR2 between proliferating T and macrophages/DC. The upregulated interactions were mainly due to *IFNG* overexpression in proliferating T cells after COVID-19 infection ( $\log_2FC = 5.1$  & adjusted  $P$ -value = 0.002, Supplementary Fig. S14). To be noted, the interaction of IFNG-IFNGR1 was detected to be significantly upregulated between proliferating T and T cells, whereas IFNG-IFNGR2 was not due to large variances of *IFNGR2* in T cells (Supplementary Fig. S15). The same phenomena were observed for IFNG-IFNGR2 in the autocrine proliferating T interactions (Supplementary Fig. S16). CellChat and iTALK, in contrast, performed differential expression at cell-level by pooling cells, resulting in biased results toward patients with large number of cells. They detected upregulation of *IFNGR1* and *IFNGR2* in macrophages, and even their downregulation in proliferating T cells, which were not true at the patient-level (Supplementary Fig. S14). The downregulation of receptors (*IFNGR1* or *IFNGR2*) led to undetermined/wrong directions of interaction changes. Another example about important CCL8-CCR5 interactions was described in the Supplementary File. In summary, the analysis strategy employed by scLR enables to find more biologically meaningful results than CellChat and iTALK. First, differential analysis at patient-level is able to find consistent changes across patients, such as *IFNG* upregulation in proliferating T cells. In contrast, differential analysis at cell-level by pooling cells together (CellChat and iTALK) was likely to be dominated by patients with large number of cells, resulting in biased results, such as false downregulation of *IFNGR1* and *IFNGR2* in proliferating T cells. Second, the strategy of considering ligands and receptors simultaneously helps remove false positives and false negatives. Although *IFNG* was strongly upregulated in proliferating T cells, scLR detected a non-significant change of IFNG-IFNGR2 between proliferating T and T cells due to large variances of *IFNGR2* in T cells, which would be a false positive if only consider *IFNG* alone. Moreover, 16% of 1313 interactions did not have significantly altered ligands or receptors even at a loose cutoff (adjusted  $P$ -value  $< 0.1$ ) (Supplementary Fig. S12), which would be missed if only consider ligands or receptors alone (false negatives). Third, scLR is able to rank dysregulated interactions by their statistical results, which is very useful to find the most important cell-type specific interactions. For example, the most dysregulated IFNG-interaction was found between proliferating T and macrophages (Supplementary Fig. S17).

### 3.3 Identification of dysregulated ligand–receptor interactions in PF and AVP

We applied scLR on two additional single-cell RNA-seq datasets generated from patients with pulmonary fibrosis (PF) (Habermann et al., 2020) and Anterior vaginal prolapse (AVP) (Li et al., 2021).

The PF data contained 12 cell types, such as type II alveolar cells (AT2), fibroblasts, endothelial cells. Using the curated ligand–receptor database (LRdb) compiled by SingleCellSignalR (Cabello-Aguilar et al., 2020), scLR identified 567 dysregulated ligand–receptor interactions in IPF patients compared to controls (adjusted  $P$ -value  $< 0.01$ ) (Supplementary Table S4). The most upregulated interactions all involve TGF- $\beta$ 1 signaling from AT2 cells, such as TGF $\beta$ 1-ITGB6, TNC-ITGAV, TGF $\beta$ 1-ITGB1 and TGF $\beta$ 1-TGFBR1/TGFBR2. TGF- $\beta$ 1 is a master regulator of ECM accumulation and a key driver of lung fibrosis (Fernandez and Eickelberg, 2012; Yue et al., 2010). Integrins  $\alpha V\beta 3$ ,  $\alpha V\beta 5$ ,  $\alpha V\beta 8$  and  $\alpha V\beta 6$  play vital roles in TGF- $\beta$  activation in fibrotic disorders (Dong et al., 2017; Fernandez and Eickelberg, 2012; Yue et al., 2010). Expression of *TNC*, an extracellular matrix protein, is significantly upregulated in fibrotic lungs, which is induced by *TGF $\beta$ 1* and contributes to TGF- $\beta$  mediated lung fibrosis (Estany et al., 2014). One recent study found that sustained elevated mechanical tension, the most common driver of lung fibrosis, activates a TGF- $\beta$  signaling loop in AT2 cells and then leads to the periphery-to-center progression of lung fibrosis (Wu et al., 2020). This is consistent with our results, which discovered strong TGF- $\beta$  activation between AT2 cells and other cell types.

The AVP dataset involved seven cell types, endothelial cells, fibroblasts, lymphatic endothelial, macrophages, myoepithelial cells (MEP), smooth muscle cells and T cells. scLR detected 34 dysregulated ligand–receptor interactions in AVP patients compared to control samples (adjusted  $P$ -value  $< 0.01$ ) (Supplementary Table S5). Endothelial cells and fibroblast participate in the highest level of changes in ligand–receptor interactions. The most upregulated interactions are mainly related to ECM organization, such as APP-LRP1 and HSPG2-LRP1 between fibroblasts and endothelial cells, autocrine A2M-LRP1 in endothelial cells, autocrine ECM1-CACHD1 in fibroblasts. LRP1 regulates ECM remodeling (Gaultier et al., 2010), while its partner A2M and APP are all involved in ECM organization. Heparan sulfate proteoglycan 2 (HSPG2), also known as perlecan, is a large multi-domain extracellular matrix proteoglycan. ECM1, extracellular matrix protein 1, is known to be upregulated in pelvic organ prolapse (Cecati et al., 2018). This agrees with previous findings that POP is an acquired disorder of extracellular matrix (Budatha et al., 2011).

## 4 Discussion

We presented scLR, a statistical method for differential analysis of ligand–receptor interactions on single-cell transcriptomics data. scLR models the distribution of the product of ligand and receptor expressions and takes the inter-sample variances, small sample sizes and dropout events into account. Overall, scLR achieved the best performance than other methods in four simulation settings, especially when the ligand and receptor are both dysregulated, or they have complementary alterations. Applied on single-cell RNA-seq datasets from severe SARS-COV-2 infection, scLR revealed the important dysregulated interactions between macrophages and proliferating T cells that lead to persistent infection. Moreover, scLR discovered activated TGF- $\beta$  signaling from alveolar type II cells contributing to the pathogenesis of pulmonary fibrosis.

scLR is designed for single-cell RNA-seq datasets with small sample sizes, which assesses sample variances by shrinkage of the estimated variances toward a pooled estimate using a Bayesian method. It can work for data with only one sample in either one of the two conditions based on the additional assumption that variances of ligand/receptor expressions in the two conditions are the same (Smyth, 2004). When there are enough samples, model-free methods, such as permutation, might be better for estimating the background distribution of expression products of ligands and receptors rather than assuming bivariate normal distributions. However, it is difficult to determine the exact number of sufficient sample sizes. The sample size depends on the number of permutations needed to generate an accurate  $P$ -value, whereas the minimum number of permutations is subject to the total number of ligand–receptor pairs and pair-wise cell types combinations.

The current version of scLR only compares ligand–receptor interactions between two conditions. If there are any batch effects, they should be removed by single-cell RNA-seq batch correction methods (Tran *et al.*, 2020) before the application of scLR. We plan to extend scLR for experiments with complex designs, which would broaden its applicability and also remove potential biases from confounders such as batch effects. Moreover, scLR only models the interactions between ligands and receptors without considering other stimulatory and inhibitory cofactors. Combining ligands, receptors and cofactors together would require new modeling framework. Identification of dysregulated LR interactions from single-cell RNA-seq alone would introduce false-positive results since cells only communicate within a certain distance. Integrating single-cell RNA-seq with spatial transcriptomics would further narrow down the important communications between cells.

## Data availability

Data from COVID-19, pulmonary fibrosis and anterior vaginal prolapse can be accessed through the Gene Expression Omnibus (GEO) with accession numbers GSE155249, GSE135893, and GSE151202 respectively.

## Funding

This work was supported by the National Cancer Institute [U2C CA233291 and U54 CA217450]; National Institutes of Health [P01 AI139449]; and Cancer Center Support Grant [P30CA068485].

*Conflict of Interest:* none declared.

## References

- Almet, A.A. *et al.* (2021) The landscape of cell–cell communication through single-cell transcriptomics. *Curr. Opin. Syst. Biol.*, **26**, 12–23.
- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Armingol, E. *et al.* (2021) Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.*, **22**, 71–88.
- Bassez, A. *et al.* (2021) A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nat. Med.*, **27**, 820–832.
- Browaeys, R. *et al.* (2020) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods*, **17**, 159–162.
- Budatha, M. *et al.* (2011) Extracellular matrix proteases contribute to progression of pelvic organ prolapse in mice and humans. *J. Clin. Invest.*, **121**, 2048–2059.
- Cabello-Aguilar, S. *et al.* (2020) SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.*, **48**, e55.
- Cecati, M. *et al.* (2018) Expression of extracellular matrix and adhesion proteins in pelvic organ prolapse. *Cell. Mol. Biol. (Noisy-le-Grand)*, **64**, 142–148.
- Cheng, X. *et al.* (2021) Single-cell analysis reveals urothelial cell heterogeneity and regenerative cues following cyclophosphamide-induced bladder injury. *Cell Death Dis.*, **12**, 446.
- Dong, X. *et al.* (2017) Force interacts with macromolecular structure in activation of TGF-beta. *Nature*, **542**, 55–59.
- Efremova, M. *et al.* (2020) CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.*, **15**, 1484–1506.
- Estany, S. *et al.* (2014) Lung fibrotic tenascin-C upregulation is associated with other extracellular matrix proteins and induced by TGFbeta1. *BMC Pulm. Med.*, **14**, 120.
- Fernandez, I.E. and Eickelberg, O. (2012) The impact of TGF-beta on lung fibrosis: from targeting to biomarkers. *Proc. Am. Thorac. Soc.*, **9**, 111–116.
- Gaultier, A. *et al.* (2010) LRP1 regulates remodeling of the extracellular matrix by fibroblasts. *Matrix Biol.*, **29**, 22–30.
- Gong, L. *et al.* (2021) Comprehensive single-cell sequencing reveals the stromal dynamics and tumor-specific characteristics in the microenvironment of nasopharyngeal carcinoma. *Nat. Commun.*, **12**, 1540.
- Grant, R.A. *et al.*; The NU SCRIPT Study Investigators. (2021) Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature*, **590**, 635–641.
- Habermann, A.C. *et al.* (2020) Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.*, **6**, eaba1972.
- Hildreth, A.D. *et al.* (2021) Single-cell sequencing of human white adipose tissue identifies new cell states in health and obesity. *Nat. Immunol.*, **22**, 639–653.
- Hou, R. *et al.* (2020) Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.*, **11**, 5011.
- Hu, Y. *et al.* (2021) CytoTalk: de novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv*, **7**, eabf1356.
- Jin, S. *et al.* (2021) Inference and analysis of cell–cell communication using CellChat. *Nat. Commun.*, **12**, 1088.
- Li, Y. *et al.* (2021) Single-cell transcriptome profiling of the vaginal wall in women with severe anterior vaginal prolapse. *Nat. Commun.*, **12**, 87.
- Martin, J.C. *et al.* (2019) Single-Cell analysis of crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell*, **178**, 1493–1508.e20.
- Noel, F. *et al.* (2021) Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.*, **12**, 1089.
- Qiu, P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1169.
- Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
- Solovey, M. and Scialdone, A. (2020) COMUNET: a tool to explore and visualize intercellular communication. *Bioinformatics*, **36**, 4296–4300.
- Speranza, E. *et al.* (2021) Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. *Sci. Transl. Med.*, **13**, eabe8146.
- Tian, Y. *et al.* (2021) Single-cell transcriptomic profiling provides insights into the toxic effects of zearalenone exposure on primordial follicle assembly. *Theranostics*, **11**, 5197–5213.
- Wu, H. *et al.* (2020) Progressive pulmonary fibrosis is caused by elevated mechanical tension on alveolar stem cells. *Cell*, **180**, 107–121.e17.
- Yue, X. *et al.* (2010) TGF-beta: titan of lung fibrogenesis. *Curr. Enzym Inhib.*, **6**, 2.
- Zhang, Y. *et al.* (2021) Cellinker: a platform of ligand–receptor interactions for intercellular communication analysis. *Bioinformatics*.