rCom: A route-based framework for inferencing inter cell type communication and regulatory network using single cell data

Honglin Wang†  
 Computer Sci. and Engr. Dept., Univ. of Connecticut, Storrs, USA  
 honglin.wang@uconn.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

ABSTRACT

The foundation of analyzing complex intercellular responses is to map ligand-receptor pairs. With recent advances of single cell RNA (scRNA) sequencing technology, and easy availability of scRNA data, several methods have been proposed to infer cell-cell communication by analyzing ligand-receptor pairs and their expressions in cell-pairs. However, these methods consider only ligand-receptor pairs to infer the communication, thereby missing information from the known portion of pathways that are upstream to ligand and downstream from receptor. In this paper, we present a novel framework, called rCom, that infers cell-cell interactions by considering the portions of pathways that are directly associated with ligand and receptors. Pathway databases like KEGG have limited annotation on ligand-receptor pairs. Therefore, the rCom framework combines knowledge from multiple biological databases such as (i) transcription factor-target database, (ii) ligand-receptor database and (iii) gene signaling pathway databases. The rCom framework extracts information by combining of these databases and form multiple communication route pairs: the communication ligand routes and communication receptor routes. A novel algorithm and a heuristic rule are used to score each route pairs between each cell groups. Finally, permutation test is used to find the significant route pairs between each two cell groups. A route based inter cell type regulatory network is generated based on the significant routes.

We demonstrate how rCom can be used to inference the inter cell type communications on two independent bone marrow datasets. Our literature survey shows the correctness of rCom in independent gene level, but the discovered communication among cell types merit wet-lab experiments for validation.

KEYWORDS

Cell communication, Topology and route-based pathway analysis, Bone marrow, Single cell RNA seq

1 Introduction

Complex intercellular responses start with binding of a ligand to its cognate receptor to activate specific cell signaling pathways. Mapping ligand- receptor pairs is fundamental to understanding how cell responds to signaling from neighboring cells and to decode the intercellular communication networks. Single cell RNA-seq technology holds great promise for studying cell-cell communication at much higher resolution. Using scRNA-seq data, several methods have been developed to infer ligand-receptor pair communications between two cell types. Skelly et al. [1] and Kumar et al. [2] predict ligand-receptor pairs if the two genes are highly expressed in the two respective cell types. Zhou et al. [3] and Vento-Tormo et al. [4] identify ligand-receptor pairs whose expressions are specific to the cell types considered. Signaling pathways are highly dynamic, and crosstalk among them is prevalent. Because of these two features, simply examining expression levels of ligand and receptor genes cannot reliably capture the overall activities of signaling pathways and interactions among them [5], [6]. As a step forward, SoptSC from Wang et al. [7] and NicheNet from Browaeys et al. [8] identify both ligand-receptor pairs and genes downstream of them. Hu et al [9] proposed a method, called CytoTalk, to generate a signal transduction network using single-cell transcriptomic data.

With the advent of topology-based pathway analysis, a particular kind, namely route-based, has shown great promise by considering gene-gene regulatory relationships in certain portions or routes of pathways. In our previous work, we analyzed TCGA dataset to demonstrate the strength of route-based pathway analysis by identifying and scoring transcription factor (TF)-centric routes in pathways [10]. Some of these pathway routes can lead to regulation of a ligand, and other routes can be directly activated by binding of the ligand. Analyzing these routes which are directly associated with ligand-receptor binding can give us better understanding of cell-cell interactions (CCIs). In this manuscript, we present a novel computational framework, named rCOM, that uses routes of biological pathways to identify the statistically significant CCIs and visualize them as a network called route-based inter cell type regulatory network (rICRN). In the rICRN, nodes represent cell groups and edges represent interaction between the nodes.

The rCom framework infered cell type communication by mapping the preprocessed scRNA-seq transcriptomic data to two types of communication routes: communication ligand route and communication receptor route which are extracted from well-established biological knowledge databases. Each route in every cell is scored by considering the consistency of regulatory pattern of gene in routes. The inter cell type communication is inferenced by a heuristic rule based on the comparison of the inter and intra cell type communication. In the end, p-value of permutation test is used to identify the statical significance for communication between two cell types through each route. A route based inter cell type regulatory network (rICRN) are built based on significant communication route pairs. The details will be introduced in section 2. We demonstrate how rCOM can be used to analyze single cell datasets in two independent bone related scRNA-seq dataset and discuss the result in the section 3. The conclusion is stated in the last section.

2 System and Methods

The overall process of the rCom framework is shown in Fig. 1. Three different databases including: TF-target database, ligand-receptor pairs database and gene signaling regulated route databases are used to build communication routes. The source databases are downloaded in text or KGML format (Fig 1a). All the nodes (ligands, receptor, gene, and TF) and edges (gene regulated relationship) are extracted from databases and stored in a graph structure. In the graph structure, nodes regulated interactions are directional and are encoded as either activation or inhibition (Fig.1a, 1c).

The input data are revealed in fig 1b. Single cell matrix contains gene expression values which are preprocessed through quality control, normalization, scaled and log transform. In the matrix, columns indicate cells and rows indicate genes. After log transformed, rCom treated gene with negative value as inhibited genes and gene with positive value as activated genes.

2.1 Communication Routes

Unlike other methods that only consider ligand-receptor, the rCom infers cell communication based on the communication routes that are extended from L-R pairs. In figure X, we show how we generate communication routes from four different biological knowledge databases including: Chip-seq [11], CellChat[12] , CellTalkDB[13] and rPAC.

Communication route pair has two routes: (i) communication ligand route (CLR) and (ii) communication receptor route (CRR). Communication ligand route captures the signals which are cell-secreted ligands produced when the transcription factor (TF) binds to the target genes. The cell that secrets ligands is known as secretor. CLR usually starts from TF and follow the path to the ligands. Two biological knowledge databased are used to generate CLR, rPAC routes and Chip-seq DB. The CLR is like effector routes in rPAC, but the primary TFs of CLR are not only from KEGG but also from Chip-seq. Communication receptor route captures the signals transduction from cell surface into the nucleus in which each route starts from receptor and follows a path to the TFs. The cell that receiving ligand signal is known as receiver. CRR are transformed from signaling routes in rPAC (KEGG). The connection of CLR and CRR are ligand and receptor pairs derived from two ligand-receptors pairs databases: CellChatDB, CellTalk and ligand-receptors pairs in p1 routes of rPAC. Totally 104004 communication route pairs are built and stored in graph structure in rCom.

2.2 Node expected value and evaluation value

*2.2.1 Node expected value.* A node expected value () is either +1 or -1 which indicates whether node should be up-regulated or down-regulated in the route to activate ligand or activated by receptor. A node is assigned an expected value using propagation method which starts from ligands in CLR or receptors in CRR to the end node of the route (TF in both CLR and CRR). The equation of this method is given in Eq 1.

(1)

In this equation, k stands for the kth nodes from the starter (ligand in CLR and receptor in CRR.). ek stands the edge between ek and ek-1. The value of ek can be assign using Eq 2.

*2.2.2 Node evaluation value.* Node evaluation value () is assigned based on the node expected value and node type. Two types of nodes will appear in the communication routes: singleton node and bundle node. A singleton node stands for only one gene in the node which is computed using Eq 3.

Diagram

Description automatically generated

Fig 2. The overview of the communication route. The blue route is CLR, and the red route is CRR. The CLR is generated from rPAC routes and Chip-seq database. The CRR is generated from rPAC routes. The ligand and receptor pairs are derived from CellChatDB, CellTablkDB and rPAC.

stands for the expression value of gene g in node k after preprocessed. Bundle node is composed of multiple nodes which collectively contribute to regulate the downstream genes. Bundle nodes are classified into AND and OR bundles based on the underlying biology. As defined in rPAC, AND bundle means that the node in the bundle should be activated simultaneously to send a strong up-regulatory signal to its downstream nodes. An AND bundle node with all activated nodes is count as activated and with both activated and inhibited nodes are count as inhibited. Eq 4 are used to assign the node value for AND bundle.

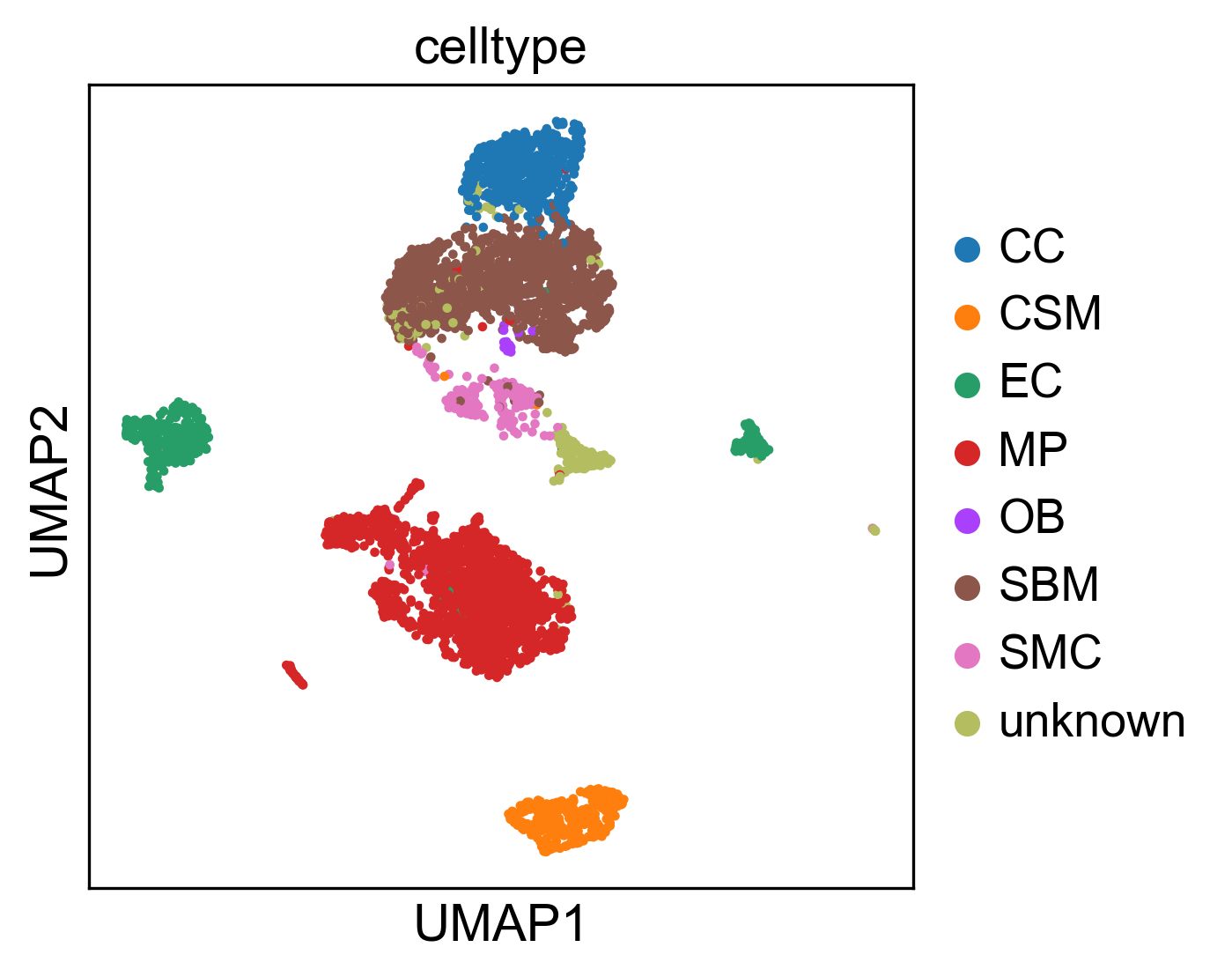
Considering the property of AND bundle, we use the node with the smallest absolute value in the bundle to represent the bundle. In contrast, OR bundle node sends its signal in two different scenarios. To send a strong up-regulatory signal to its downstream, only one of the nodes in the bundle needs to be activated. But to send a down-regulatory signal, all nodes in the bundle should be inhibited. An OR bundle node will be counted as activated if one or more nodes in the bundle is activated and with all nodes are inhibited in the bundle, it will be counted as inhibited. Eq. 5 are generated to assign the node value for OR bundle. By considering the property of OR bundle, we use the node with largest absolute value to represent the entire bundle.

2.3 Communication route score

A communication route () score () measure the capability of cell to communicated with other cells through communication route . A cell with high CLR score means it has a high capability to secret ligand which are highly regulated in this specific route and with a high CRR score indicates a high capability to receive signals through the receptor and regulate it downstream gene and TF. The score is assigned using the equation given in Eq. 6a, 6b.

where is the number of nodes in communication route , and are the node evaluation value of ligand node or receptor node in CLR or CRR. Since the communication behavior is mostly rely on the ligand and receptor, an adjustable hyperparameter is given to emphasize the effect of ligand or receptor. The choice of will be “context-dependent” rang from 0 to 1 and need to be determined empirically like hyperparameters of machine learning models.

2.4 Inference of intercellular communications

Fig 3. The reproduce 2-D visualization of UMAP, colored by cell label provided by studies.

To infer the inter and intracellular communications, a novel heuristic rule are invented based on the two distinct types of cell-cell communication: autocrine and paracrine. Since the mass of ligand secreted by a cell type is constant, a strong inter cell type communication will lead to a week intra cell type communication and vice versa. According to this rule, a score of communication between the cell type and through communication route pair are computed using equation given in Eq.7.

Here and represent the CLR and CRR score in cell group and were computed using Eq. 8a and 8b. and stands for the score of communication route for cell in cell group .

By checking the communication score , it is possible to inference the cell communication between the whole dataset. The higher score represents for a higher probability that communication between the cell in two cell type through the specific communication routes.

2.5 Identification of statistically significant intercellular communication routes

To prove the inference of intercellular communication is not coincidence, the identification of statistically significance is necessary. The significant interactions between two cell groups are identified using permutation test by randomly permuting the cell type label and then recalculating the communication score between cell type and through communication route pair . The p-value of each communication score is calculated using Eq. 9

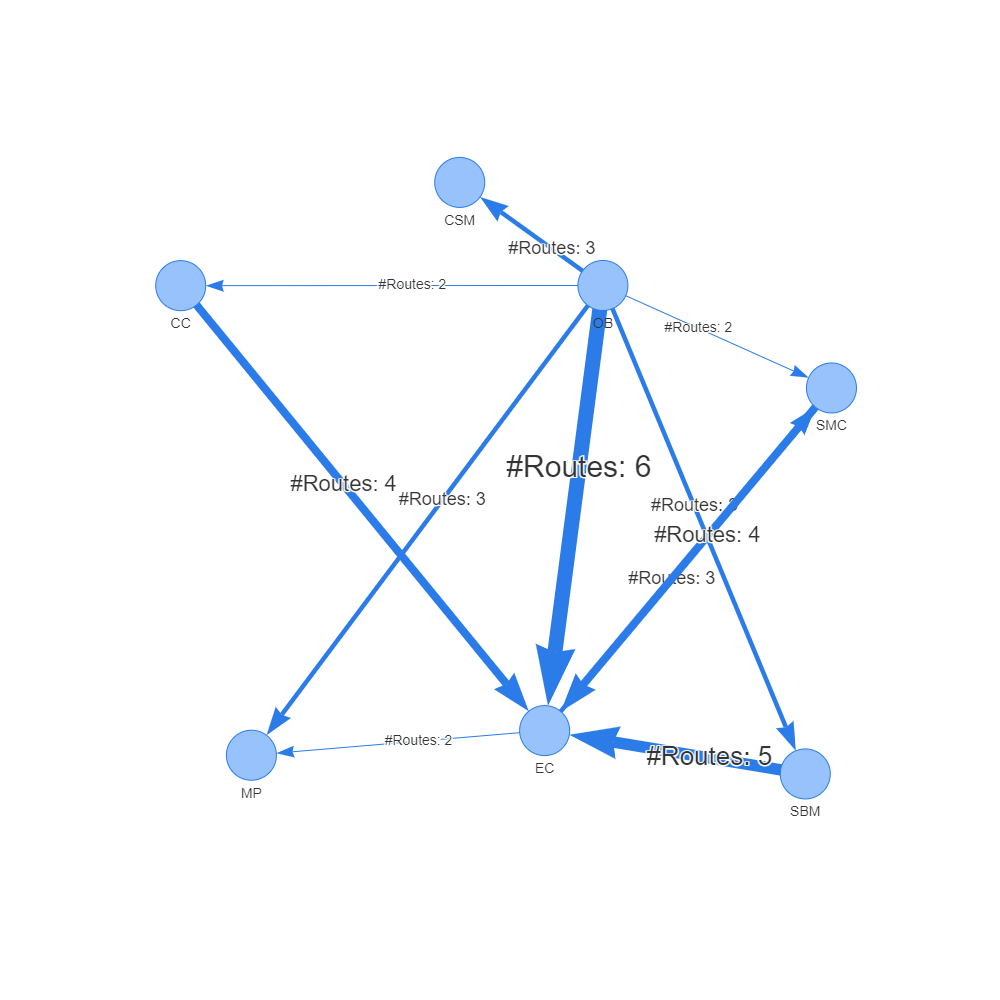


Fig. 4 rICRN generated from rCOM, the weight of edge shows the number of communication routes between two cell group higher than the threshold.

Where the score is the communication probability for the m-th permutation. M is the total number permutations (default is 100). The communications with p-value < 0.05 are considered as significant.

3 Experiment results

To demonstrate the correctness and practicality of rCom. We show how independent two single cell datasets can be analyzed using rCom with 4 biological knowledge databases as mentioned in methods part. We built 2 rICRNs based on the significant communication route pairs and validate the result by checking the gene expression value of ligand and receptor and by literature survey.

3.1 Experiments with Acta2-lineage cells in osseointegration dataset

The first study that applied to rCom identified the specific bone marrow stromal cell populations that contribute to bone formation around metal implants [14]. After preprocess, the scRNA-seq data matrix we collected from this study has 4397 cells and 31053 features (genes). The labels of each cell group are given in this dataset including: cell of skeletal muscle (CSM) (269 cells), chondrocyte (CC) (545 cells), endothelia cell (EC) (394 cells), macrophage (MP) (1608 cells), osteoblast (OB) (19 cells), smooth muscle cell (SMC) (191 cells) and stromal cell of bone marrow (SBM) (1371 cells). To prove the dataset that applied in the rCom are correctly preprocessed, the 2-D visualization of UMAP with the given labels is shown in the Fig 3.

The significant communication routes of intercellular communication are shown in Tab. 1 with the threshold of the score larger than 8. And the rICRN of this dataset using the highest score route is shown in Fig 4 generated using pyvis. By checking the networks, we notice that the communication from SBM to EC, from OB to EC and from CC to EC have the most significa nt communication routes. To prove the correctness of the inferred communications, we generated the violin plot with the ligand and receptor gene expression values in the datasets. In Fig. 5, we try to validate the identified communication route (id: 70072) between EC and MP through the ICAM2 – ITGB2 pairs. The gene expression highly supports our results with the ICAM2 highly expressed in the EC group and poorly expressed in the MP group while on the receptor ends, ITGB2 is highly expressed in the MP cells but poorly expressed in the EC. Our literature survey shows that in 2011, Zhang et al. [15] identified the communication between EC and MP through ICAM family gene. In another case, OB are known to communicate to MP cell through VEGFA as state by Mayer et al. [16]. We identified the communication happens between these two cell types through communication route with id 103917. The violin plot of the ligand and receptor in this route proves our result as shown in figure. The VEGFA gene are relatively highly expressed in the OB group and poorly expressed in the EC group. But the as receptor of VEGFA, KDR is highly expressed in the EC and poorly expressed in the rest of cell groups.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S** | **R** | **R id** | **Score** | **Ligand** | **Receptor** | **Pval** |
| OB | EC | 82280 | 9.83 | IBSP | ITGB3 | <0.0005 |
| OB | MP | 96177 | 9.53 | IBSP | ITGAV | <0.0005 |
| OB | SMC | 96177 | 9.18 | IBSP | ITGAV | <0.0005 |
| OB | SBM | 96177 | 9.05 | IBSP | ITGAV | <0.0005 |
| *EC* | *MP* | *70072* | *9.04* | *ICAM2* | *ITGB2* | *<0.0005* |
| OB | EC | 6452 | 8.68 | TNC | PTPRB | <0.0005 |
| EC | OB | 77882 | 8.66 | GNAI2 | LPAR3 | <0.0005 |
| OB | EC | 103917 | 8.65 | VEGFA | KDR | <0.0005 |
| OB | EC | 4562 | 8.64 | SEMA5A | MET | <0.0005 |
| OB | EC | 103718 | 8.63 | PDGFC | KDR | <0.0005 |
| OB | EC | 89677 | 8.55 | IBSP | ITGAV | <0.0005 |
| SMC | EC | 103686 | 8.53 | VTN | KDR | <0.0005 |
| OB | SMC | 95280 | 8.49 | IBSP | ITGB3 | <0.0005 |
| OB | CSM | 93655 | 8.48 | IBSP | ITGB3 | <0.0005 |
| OB | CSM | 94552 | 8.46 | IBSP | ITGAV | <0.0005 |
| OB | SBM | 82280 | 8.42 | IBSP | ITGB3 | <0.0005 |
| CC | EC | 103767 | 8.42 | GREM1 | KDR | <0.0005 |
| SBM | EC | 103695 | 8.41 | PDGFC | KDR | <0.0005 |
| OB | CC | 94552 | 8.37 | IBSP | ITGAV | <0.0005 |
| EC | SMC | 11106 | 8.29 | GNAI2 | EDNRA | <0.0005 |
| SBM | EC | 6448 | 8.28 | TNC | PTPRB | <0.0005 |
| EC | SMC | 81908 | 8.27 | JAG2 | NOTCH3 | <0.0005 |
| CC | EC | 6448 | 8.24 | TNC | PTPRB | <0.0005 |
| EC | MP | 96764 | 8.24 | COL4A1 | ITGAV | <0.0005 |
| SBM | EC | 6478 | 8.20 | PTN | PTPRB | <0.0005 |
| EC | SMC | 81697 | 8.20 | DLL4 | NOTCH3 | <0.0005 |
| OB | MP | 95280 | 8.19 | IBSP | ITGB3 | <0.0005 |
| OB | SBM | 4562 | 8.19 | SEMA5A | MET | <0.0005 |
| CC | EC | 103929 | 8.17 | VEGFA | KDR | <0.0005 |
| SMC | EC | 6457 | 8.17 | TNC | PTPRB | <0.0005 |
| SMC | EC | 103693 | 8.15 | PDGFC | KDR | <0.0005 |
| CC | EC | 82241 | 8.14 | COMP | ITGB3 | <0.0005 |
| SBM | EC | 103664 | 8.12 | VEGFA | KDR | <0.0005 |
| OB | CC | 87155 | 8.10 | IBSP | ITGB3 | <0.0005 |
| OB | MP | 4562 | 8.10 | SEMA5A | MET | <0.0005 |
| SBM | EC | 103769 | 8.09 | VEGFD | KDR | <0.0005 |
| OB | CSM | 4562 | 8.07 | SEMA5A | MET | <0.0005 |
| SMC | EC | 103664 | 8.02 | VEGFA | KDR | <0.0005 |
| CSM | EC | 103769 | 8.01 | VEGFD | KDR | <0.0005 |

Table 1. The significant communication routes pairs between each two cell types with score higher than threshold (8.0). In the header, S stands for secretor and R stands for receiver.

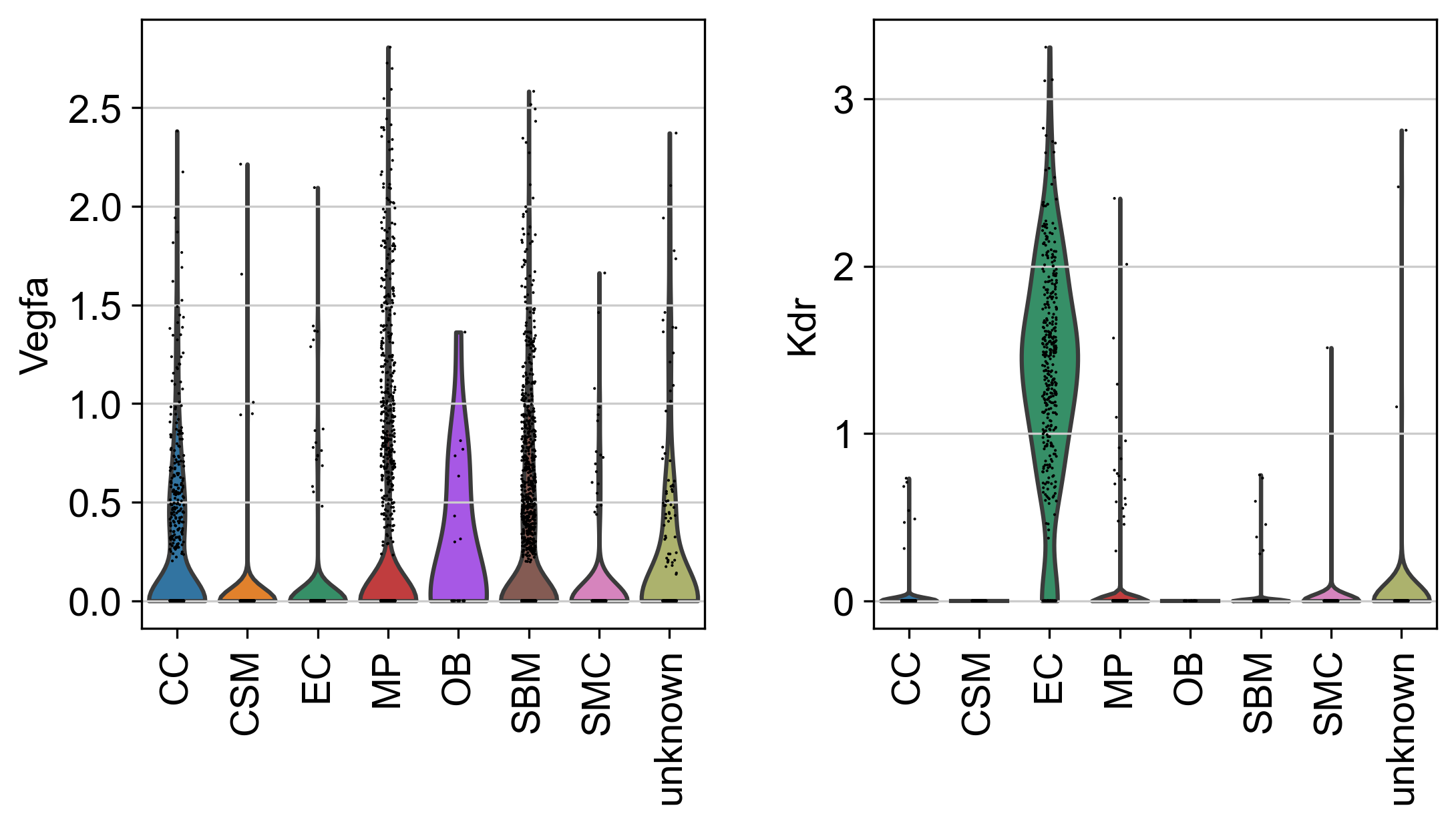


Fig 6. The violin plot of VEGFA and KDR gene expression value in each cell type.



Fig 5. The violin plot of ICAM2 and ITGB2 gene expression value in each cell type.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S** | **R** | **R id** | **Score** | **Ligand** | **Receptor** | **pval** |
| SMC | PL | 95696 | 12.64 | COL18A1 | ITGB3 | <0.0005 |
| SMC | MC | 80134 | 11.58 | LGALS1 | PTPRC | <0.0005 |
| SMC | PL | 95739 | 11.03 | COL4A2 | ITGB3 | <0.0005 |
| SMC | PL | 95421 | 10.34 | MFGE8 | ITGB3 | <0.0005 |
| SMC | PL | 95367 | 9.7 | NID1 | ITGB3 | <0.0005 |
| MC | SMC | 72456 | 9.63 | IL1B | IL1R1 | <0.0005 |
| OB | PL | 82280 | 9.6 | IBSP | ITGB3 | <0.0005 |
| OB | PL | 82619 | 9.45 | FN1 | ITGB3 | <0.0005 |
| MC | EC | 72456 | 9.44 | IL1B | IL1R1 | <0.0005 |
| CC | PL | 90899 | 9.42 | TGM2 | ITGB3 | <0.0005 |
| OB | PL | 82963 | 9.41 | CCN1 | ITGB3 | <0.0005 |
| SMC | PL | 95503 | 9.39 | FBN1 | ITGB3 | <0.0005 |
| OB | MC | 69430 | 9.36 | CCN1 | ITGB2 | <0.0005 |
| SMC | PL | 95268 | 9.35 | VTN | ITGB3 | <0.0005 |
| SMC | MC | 69220 | 9.3 | JAM3 | ITGB2 | <0.0005 |
| MC | CC | 72456 | 9.11 | IL1B | IL1R1 | <0.0005 |
| SMC | PL | 95290 | 9.08 | TGFB3 | ITGB3 | <0.0005 |
| SMC | PL | 96031 | 9.08 | ITGAV | ITGB3 | <0.0005 |
| MC | OB | 72456 | 9.04 | IL1B | IL1R1 | <0.0005 |
| OB | PL | 82885 | 9.04 | SPP1 | ITGB3 | <0.0005 |
| SMC | MC | 69430 | 9 | CCN1 | ITGB2 | <0.0005 |
| SMC | PL | 95963 | 9 | CCN1 | ITGB3 | <0.0005 |
| SMC | MC | 69337 | 8.98 | CCN2 | ITGB2 | <0.0005 |
| HP | SMC | 6394 | 8.88 | CDH1 | PTPRF | <0.0005 |
| MC | PL | 73780 | 8.88 | IL1B | IL1R1 | <0.0005 |
| MC | PL | 95780 | 8.88 | TGM2 | ITGB3 | <0.0005 |
| SMC | MC | 69032 | 8.83 | PLAT | ITGB2 | <0.0005 |
| MC | SMC | 11108 | 8.76 | GNAI2 | EDNRA | <0.0005 |
| MC | PH | 72456 | 8.75 | IL1B | IL1R1 | <0.0005 |
| OB | PL | 82724 | 8.72 | ADAM15 | ITGB3 | <0.0005 |
| OB | MC | 69220 | 8.71 | JAM3 | ITGB2 | <0.0005 |
| SMC | PL | 95841 | 8.71 | THBS2 | ITGB3 | <0.0005 |
| SMC | PL | 95640 | 8.7 | FN1 | ITGB3 | <0.0005 |
| SMC | MC | 69438 | 8.66 | SPON2 | ITGB2 | <0.0005 |
| MC | PL | 78576 | 8.61 | GNAI2 | F2R | <0.0005 |
| OB | MC | 69439 | 8.6 | SPON2 | ITGB2 | <0.0005 |
| SMC | MC | 69190 | 8.6 | VCAM1 | ITGB2 | <0.0005 |
| OB | PL | 83036 | 8.58 | ITGAV | ITGB3 | <0.0005 |
| OB | PL | 82822 | 8.57 | THBS2 | ITGB3 | <0.0005 |
| OB | PL | 83001 | 8.55 | VEGFA | ITGB3 | <0.0005 |
| HP | PL | 95712 | 8.54 | ADAM15 | ITGB3 | <0.0005 |
| MC | EC | 81406 | 8.54 | SLPI | CD4 | <0.0005 |
| MC | PH | 81406 | 8.54 | SLPI | CD4 | <0.0005 |
| MC | OB | 77886 | 8.54 | GNAI2 | LPAR3 | <0.0005 |
| MC | SMC | 81406 | 8.53 | SLPI | CD4 | <0.0005 |
| MC | OB | 81406 | 8.52 | SLPI | CD4 | <0.0005 |
| HP | PL | 95592 | 8.47 | ICAM4 | ITGB3 | <0.0005 |
| PH | MC | 70058 | 8.46 | SPON2 | ITGB2 | <0.0005 |
| SMC | MC | 69406 | 8.43 | PLAU | ITGB2 | <0.0005 |
| MC | PL | 95964 | 8.42 | VEGFA | ITGB3 | <0.0005 |
| HP | EC | 6357 | 8.39 | CDH1 | PTPRM | <0.0005 |
| OB | HP | 70050 | 8.39 | CCN1 | ITGB2 | <0.0005 |
| OB | MC | 69337 | 8.36 | CCN2 | ITGB2 | <0.0005 |
| HP | PL | 95866 | 8.35 | ITGB3BP | ITGB3 | <0.0005 |
| SMC | MC | 69175 | 8.35 | KNG1 | ITGB2 | <0.0005 |
| OB | MC | 69010 | 8.33 | PLAT | ITGB2 | <0.0005 |
| MC | SMC | 101117 | 8.32 | S100A8 | TLR4 | <0.0005 |
| MC | EC | 100813 | 8.3 | S100A8 | TLR4 | <0.0005 |

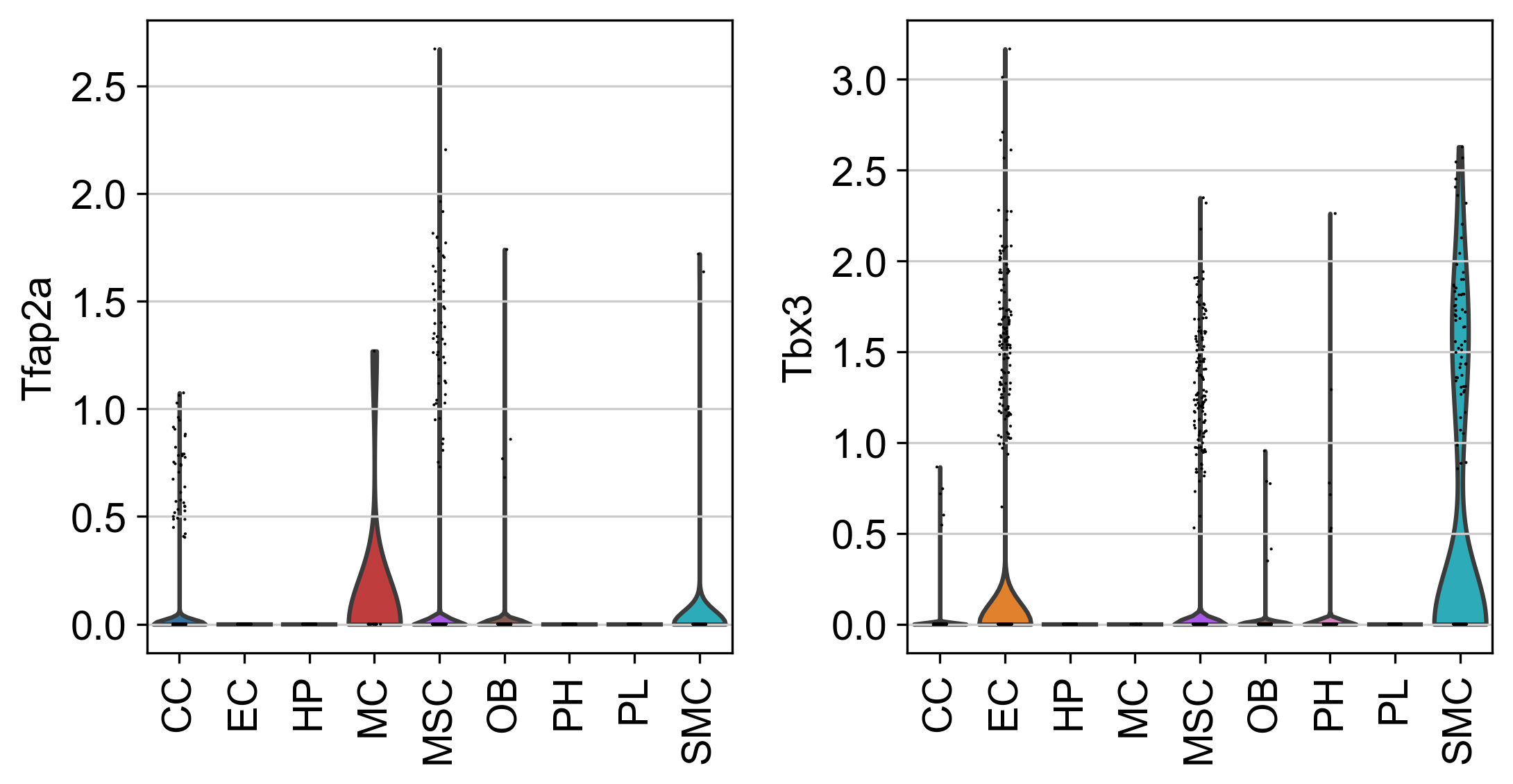
Table 2. The significant communication routes pairs between each two cell types with score higher than threshold (8.3).

A picture containing map

Description automatically generated

Fig 7. The reproduce 2-D visualization of UMAP, colored by cell label provided by studies.

3.2 Experiments with bone marrow stromal scRNA-seq of 11 weeks old mice dataset



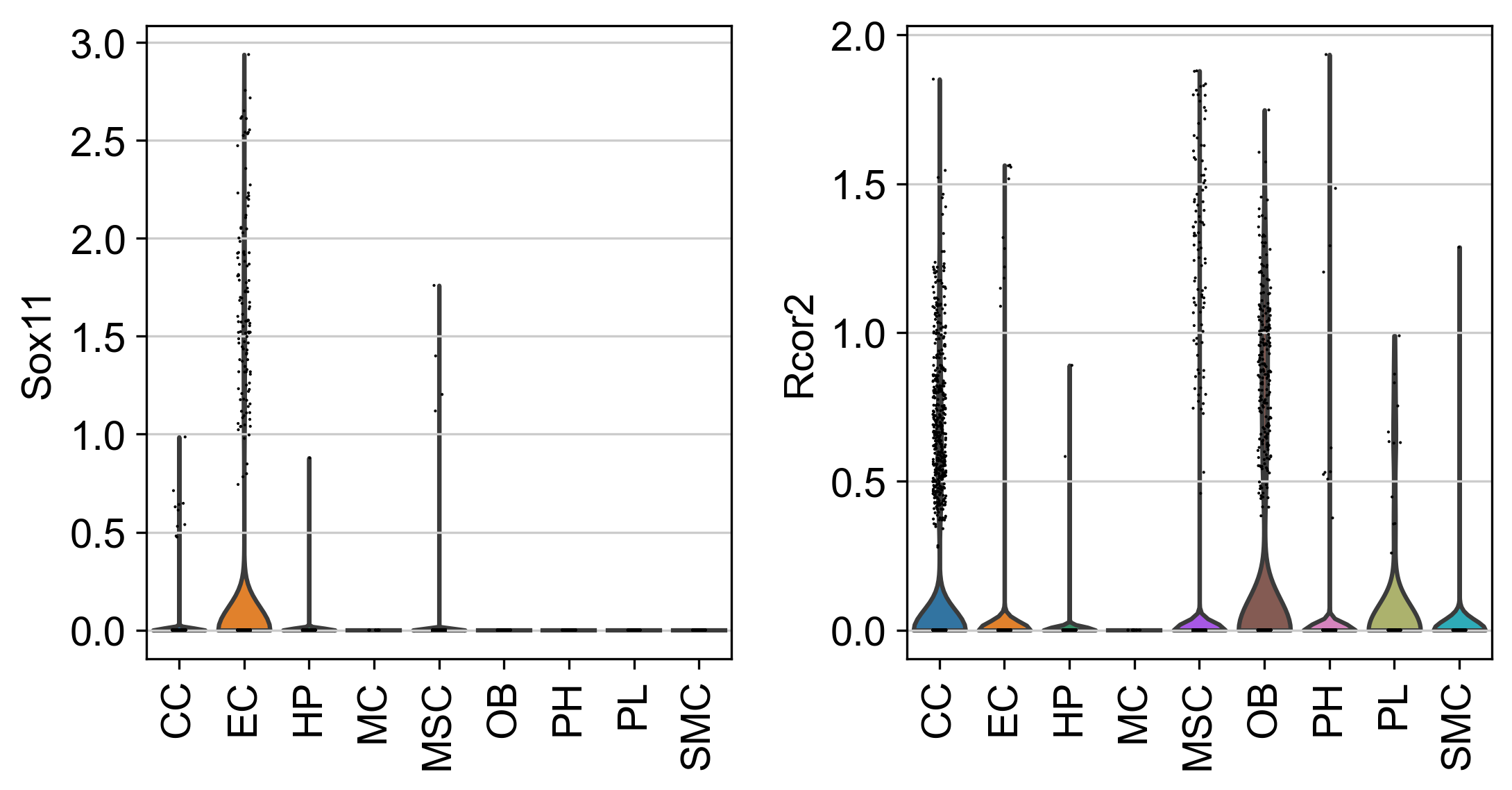


Fig 11. The violin plot for upper stream gene expression value of COL18A1 and ITGB3.

Another dataset we collected from a study [17] which explored osteogenesis in adult mice and show that a bone-targeting, high affinity without causing adverse effects in other organs, which are known to reply on intact Notch signaling interactions. After preprocess, the scRNA-seq data matrix has 17015 cells and 15046 features(genes). The labels of each cell groups are given in the meta file of the dataset including, CC (2270 cells), EC (1677 cells), Hematopoietic (HP) (849 cells), MSC (9168 cells), OB (971 cells), platelets (PL) (232 cells), proliferating Hem (PH) (1532 cells), SMC (238 cells), MC (78 cells). Our reproduced single cell 2-D plot visualization using UMAP is shown in Fig 7. The pattern those cells sharing same group are close to each other stands for the correctness of our preprocess.

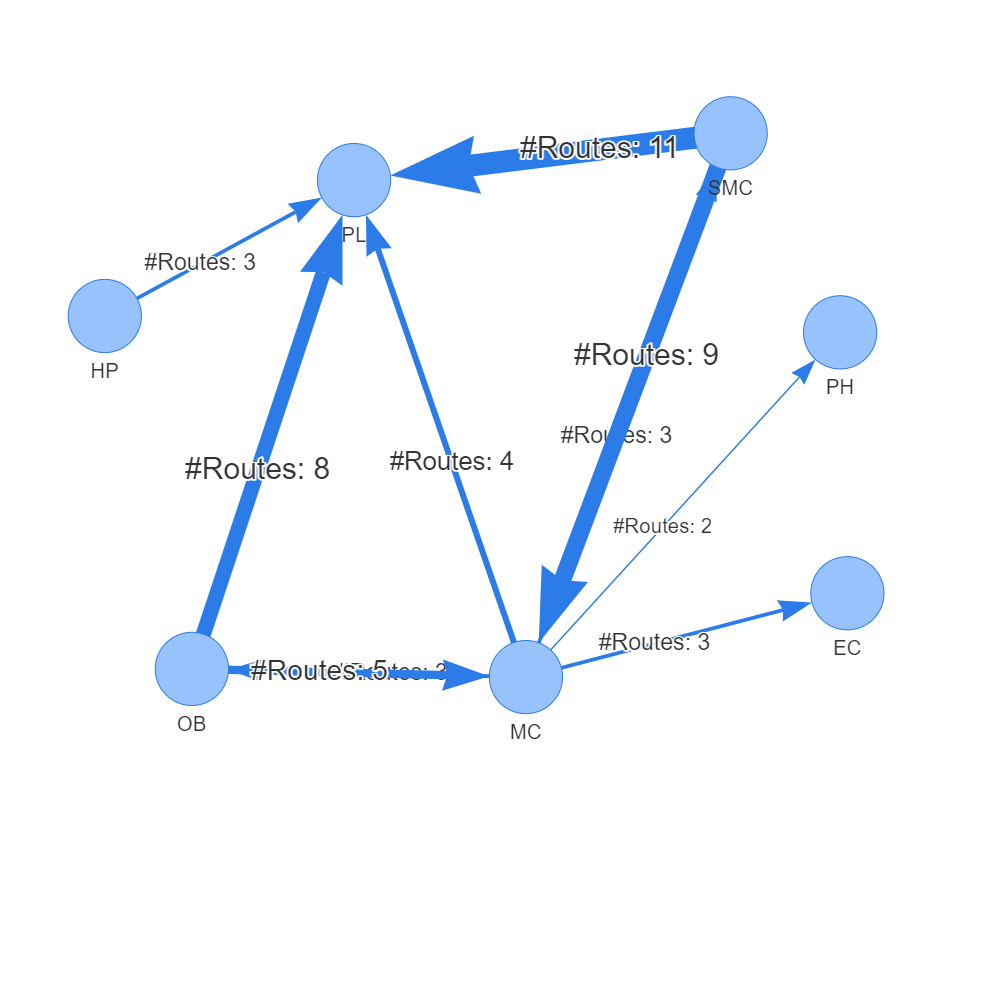


Fig. 8 rICRN generated from rCOM, the weight of edge shows the number of communication routes between two cell group higher than the threshold.

The significant communication routes identified by rCom are shown in Tab.1 with threshold of score larger than 8.3 and the rICRN are build and shown in figure 8. Among all the communication between each two cell types. The communication between SMC and PL has the largest number of significant routes pairs, and the communication from MC to PH only have 2 route pairs. To prove the correctness of the identified significant communication routes, the violin plot is used to find the gene expression of ligand and receptor in each cell groups. In Fig 9, the gene expression value of COL18A1 and ITGB3 ligand and receptor pairs are compared using violin plot. In all cell groups, COL18A1 only highly expressed in the SMC and its receptor, ITGB3 are highly expressed in the PL and only few MC cells have ITGB3 expressed. This inferred that SMC cells are communicated with PL cells through COL18A1 and ITGB3. Our literature shows that in 2018, Misra et al. [18] indicates that ITGB3 plays and important role for smoothing muscle-derived atherosclerotic plaque cells including: SMC and PL. Another case we use is the communication between SMC and MC through LGALS1-PTPRC pairs. As shown in Fig 10, LGALS1 is highly expressed in SMC, OB and HP. But its receptor, PTPRC, are only expressed in HP and MC. Which also proves rCOM’s finds in gene expression level.

Chart, diagram

Description automatically generated

Fig 9. The violin plot of COL18A1 and ITGB3 gene expression value in each cell type.

3.3 Results discussion

By checking the results of applying rCOM to two bone related datasets, we find several noticeable points as follow:

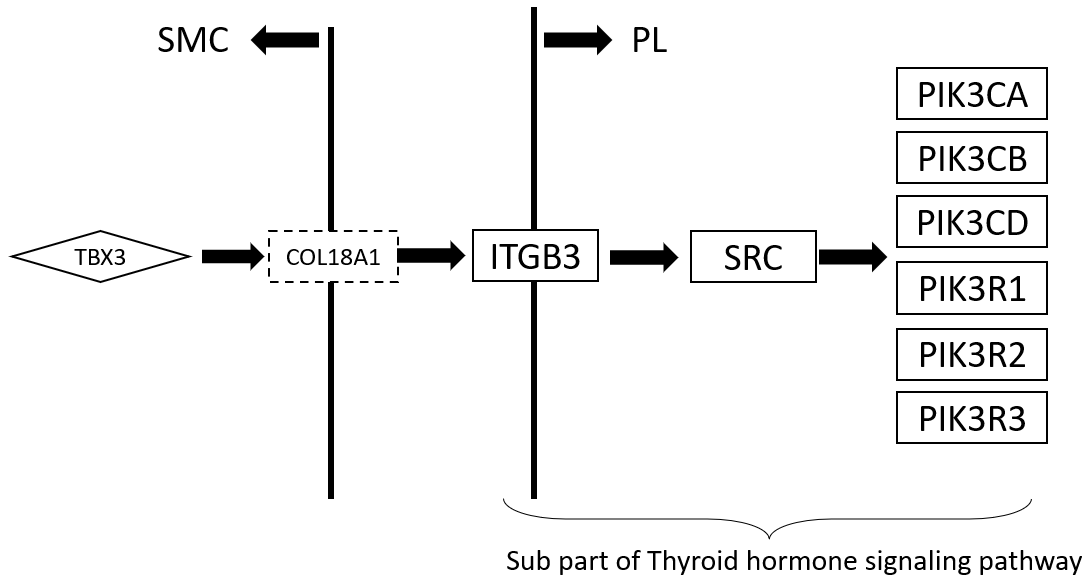


Fig 12. The complete view of communication route pairs id 95696

Chart, histogram

Description automatically generated

Fig 10. The violin plot of LGALS1 and PTPRC gene expression value in each cell type.

3.3.1 rCom not only find the ligand and receptor pairs but also identified the gene signaling route that actives the ligand genes and the routes that are activated by the receptors.In the experiments with bone marrow stromal dataset, the route pairs we find between SMC and PL through COL18A1 and ITGB3 has multiple upper TFs which are known to active COL18A1 include: TFAP2A, TBX3, NUCKS1, SOX11 and RCOR2. rCOM picks the TBX3 as the upper stream for the COL18A1 instead of other. In Fig.11, we show the gene expression of these TFs in different cell types. TBX3 are only highly expressed in the SMC. The result of this violin plot proves rCOM’s identification of upper stream is correct. The complete route is shown in Fig 12, which the TBX2 transcripts COL18A1 and actives a route in thyroid hormone signaling pathways through ITGB3.

*3.3.2 rCOM provide a new way for exploring of pathway crosstalk*. Identifying crosstalk between interacting pathways has been an active research area. Nodes-in-Common (NIC) predicts crosstalk by checking if two pathways share any proteins [19]–[21]. Hsu et.al. estimate pathway crosstalk based on similarities in Gene Ontology annotations[22]. Given a protein-protein interaction network, Li et al. links two pathways A and B if more edges connect the proteins in A to the proteins in B than expected by chance in a randomly wired network [23]. In our previous work, ctBuilder [24] is developed to identify a subnetwork interconnecting two pathways. However, rCom provide a novel aspect of exploring and identifying pathway talk, which is the ligand and receptor pairs. By analyzing the communication route pairs, unknown crosstalk between two independent pathways might be identified.

4 Conclusion

Developing rCom was inspired by

Extending

rCom combines 3 different categories of biological knowledge databases including TF-target databases, ligand-receptor databases and gene signaling pathway databases.

Among all these methods, Ctyotalk returns a signal transduction network between cell types which is similar to rCOM but the network in Cytotalk is inferred from a PCSF problem and rCOM try to identify the crosstalk between routes extracted from well-established pathway databases. To our best knowledge, no method currently exists inferring intercellular communication considering the gene regulatory patterns (activated or inhibited) and the consistence of regulatory patterns in the pathway routes at the scRNA-seq level.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] D. A. Skelly *et al.*, “Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart,” *Cell Rep.*, vol. 22, no. 3, pp. 600–610, Jan. 2018, doi: 10.1016/J.CELREP.2017.12.072.

[2] M. P. Kumar *et al.*, “Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics,” *Cell Rep.*, vol. 25, no. 6, pp. 1458-1468.e4, Nov. 2018, doi: 10.1016/J.CELREP.2018.10.047.

[3] J. X. Zhou, R. Taramelli, E. Pedrini, T. Knijnenburg, and S. Huang, “Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes,” *Sci. Reports 2017 71*, vol. 7, no. 1, pp. 1–15, Aug. 2017, doi: 10.1038/s41598-017-09307-w.

[4] R. Vento-Tormo *et al.*, “Single-cell reconstruction of the early maternal–fetal interface in humans,” *Nat. 2018 5637731*, vol. 563, no. 7731, pp. 347–353, Nov. 2018, doi: 10.1038/s41586-018-0698-6.

[5] Y. E. Antebi, N. Nandagopal, and M. B. Elowitz, “An operational view of intercellular signaling pathways,” *Curr. Opin. Syst. Biol.*, vol. 1, pp. 16–24, Feb. 2017, doi: 10.1016/J.COISB.2016.12.003.

[6] M. Billmann, V. Chaudhary, M. F. ElMaghraby, B. Fischer, and M. Boutros, “Widespread Rewiring of Genetic Networks upon Cancer Signaling Pathway Activation,” *Cell Syst.*, vol. 6, no. 1, pp. 52-64.e4, Jan. 2018, doi: 10.1016/J.CELS.2017.10.015.

[7] S. Wang, M. Karikomi, A. L. Maclean, and Q. Nie, “Cell lineage and communication network inference via optimization for single-cell transcriptomics,” *Nucleic Acids Res.*, vol. 47, no. 11, pp. e66–e66, Jun. 2019, doi: 10.1093/NAR/GKZ204.

[8] R. Browaeys, W. Saelens, and Y. Saeys, “NicheNet: modeling intercellular communication by linking ligands to target genes,” *Nat. Methods 2019 172*, vol. 17, no. 2, pp. 159–162, Dec. 2019, doi: 10.1038/s41592-019-0667-5.

[9] Y. Hu, T. Peng, L. Gao, and K. Tan, “CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data,” *Sci. Adv.*, vol. 7, no. 16, Apr. 2021, doi: 10.1126/SCIADV.ABF1356/SUPPL\_FILE/ABF1356\_TABLE\_S6.XLSX.

[10] P. Joshi, B. Basso, H. Wang, S. H. Hong, C. Giardina, and D. G. Shin, “rPAC: Route based pathway analysis for cohorts of gene expression data sets,” *Methods*, Oct. 2021, doi: 10.1016/J.YMETH.2021.10.002.

[11] “Chromatin Immunoprecipitation Sequencing (ChIP-Seq).” https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html (accessed May 11, 2022).

[12] S. Jin *et al.*, “Inference and analysis of cell-cell communication using CellChat,” *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–20, Feb. 2021, doi: 10.1038/s41467-021-21246-9.

[13] X. Shao, J. Liao, C. Li, X. Lu, J. Cheng, and X. Fan, “CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice,” *Brief. Bioinform.*, vol. 22, no. 4, Jul. 2021, doi: 10.1093/BIB/BBAA269.

[14] A. Vesprey *et al.*, “Tmem100- and Acta2-Lineage Cells Contribute to Implant Osseointegration in a Mouse Model,” *J. Bone Miner. Res.*, vol. 36, no. 5, pp. 1000–1011, May 2021, doi: 10.1002/JBMR.4264.

[15] J. Zhang *et al.*, “Regulation of Endothelial Cell Adhesion Molecule Expression by Mast Cells, Macrophages, and Neutrophils,” *PLoS One*, vol. 6, no. 1, 2011, doi: 10.1371/JOURNAL.PONE.0014525.

[16] H. Mayer, H. Bertram, W. Lindenmaier, T. Korff, H. Weber, and H. Weich, “Vascular endothelial growth factor (VEGF-A) expression in human mesenchymal stem cells: autocrine and paracrine role on osteoblastic and endothelial differentiation,” *J. Cell. Biochem.*, vol. 95, no. 4, pp. 827–839, Jul. 2005, doi: 10.1002/JCB.20462.

[17] C. Xu *et al.*, “Induction of osteogenesis by bone-targeted Notch activation,” *Elife*, vol. 11, Feb. 2022, doi: 10.7554/ELIFE.60183.

[18] F. Hartmann *et al.*, “SMC-Derived Hyaluronan Modulates Vascular SMC Phenotype in Murine Atherosclerosis,” *Circ. Res.*, vol. 129, no. 11, pp. 992–1005, Nov. 2021, doi: 10.1161/CIRCRESAHA.120.318479.

[19] D. M *et al.*, “Analysis and correction of crosstalk effects in pathway analysis,” *Genome Res.*, vol. 23, no. 11, pp. 1885–1893, Nov. 2013, doi: 10.1101/GR.153551.112.

[20] K. H and K. MR, “Abiotic stress signalling pathways: specificity and cross-talk,” *Trends Plant Sci.*, vol. 6, no. 6, pp. 262–267, 2001, doi: 10.1016/S1360-1385(01)01946-X.

[21] T. CM, E. B, and K. CR, “Critical nodes in signalling pathways: insights into insulin action,” *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 2, pp. 85–96, Feb. 2006, doi: 10.1038/NRM1837.

[22] C.-L. Hsu and U.-C. Yang, “Discovering pathway cross-talks based on functional relations between pathways,” *BMC Genomics 2012 137*, vol. 13, no. 7, pp. 1–15, Dec. 2012, doi: 10.1186/1471-2164-13-S7-S25.

[23] L. Y, A. P, and R. D, “A global pathway crosstalk network,” *Bioinformatics*, vol. 24, no. 12, pp. 1442–1447, Jun. 2008, doi: 10.1093/BIOINFORMATICS/BTN200.

[24] H. Wang, P. Joshi, S. H. Hong, D. J. Shin, and D. G. Shin, “CtBuilder: A framework for building pathway crosstalks by combining single cell data with bulk cell data,” *Proc. - 2021 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2021*, pp. 270–273, 2021, doi: 10.1109/BIBM52615.2021.9669811.