rCom: A route-based framework inferring 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

Peter F. Maye  
 Dept. of Reconstructive Sciences, UCHC, Farmington, USA pmaye@uchc.edu

Pujan Joshi  
 Computer Sci. and Engr. Dept., Univ. of Connecticut, Storrs, USA  
 pujan.joshi@uconn.edu

David W. Rowe  
 Dept. of Reconstructive Sciences, UCHC, Farmington, USA drowe@uchc.edu

Chenyu Zhang  
 Computer Sci. and Engr. Dept., Univ. of Connecticut, Storrs, USA [chenyu.zhang@uconn.edu](mailto:chenyu.zhang@uconn.edu)

Dong-Guk Shin  
 Computer Sci. and Engr. Dept., Univ. of Connecticut, Storrs, USA  
 dong.shin@uconn.edu

ABSTRACT

With recent advances of single cell RNA (scRNA) sequencing technology, several methods have been proposed to infer cell-cell communication by analyzing ligand-receptor pairs. However, existing methods have limited ways of using what we call “prior knowledge”, i.e., what are already known (albeit incompletely) about the upstream for the ligand and the downstream for the receptor. In this paper, we present a novel framework, called rCom, capable of inferring cell-cell interactions by considering portions of pathways that would be associated with upstream of the ligand and downstream of receptors under examination. The rCom framework integrates knowledge from multiple biological databases including transcription factor-target database, ligand-receptor database and publicly available curated signaling pathway databases. We combine both algorithmic methods and heuristic rules to score how each putative ligand-receptor pair may matchup between all possible cell subtype pairs. Permutation test is performed to rank the hypothesized cell-cell communication routes. We performed a case study using single cell transcriptomic data from bone biology. Our literature survey suggests that rCom could be effective in discovering novel cell-cell communication relationships that have been only partially known in the field.

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 cells respond to signaling from neighboring cells. Single cell RNA-seq technology holds great promise for studying cell-cell communication which was not easy when gene expression data is obtained from cell population-based technologies such as microarray and bulk RNA-seq. Using scRNA-seq data, several methods have been developed to infer ligand-receptor pair communications between interacting cell types. Skelly et al. [1] and Kumar et al. [2] predict ligand-receptor pairs by studying if the two genes are highly co-expressed in different cell types. More recently, the cell-cell interaction analysis has gone beyond identifying ligand-receptor pairs to produce entire gene regulation networks that could model interaction relationships among subpopulations of cells. Hu et al [3] present that CytoTalk can de novo generate a signal transduction network between a pair of cell types, which is basically interconnecting a pair of GRNs, each constructed for the cell types under consideration.

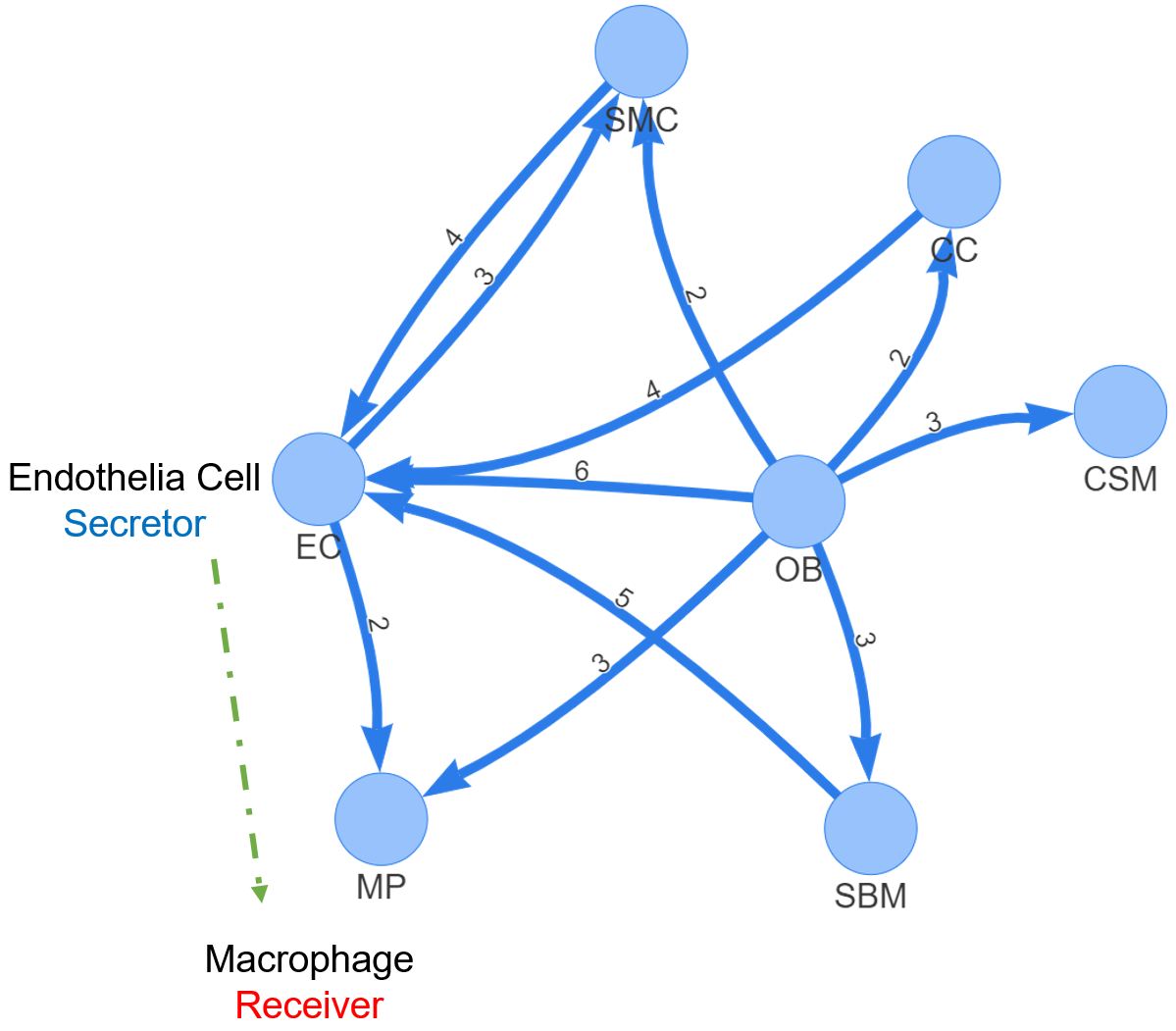
In rCom, inferring cell-cell interactions is decomposed into three parts, analysis of the upstream of the ligand, analysis of the downstream of receptors, and analysis of the most plausible ligand-receptor pairing. Our approach assumes subtyping cells are performed a priori and the labels for each cell subtype are known. Thanks to this refinement, rCom takes advantage of estimating if a certain route (branch) of a curated pathway is activated/ inhibited which in turn provides clues for picking which ligand-receptor pairs are more likely utilized in the regulatory system.

The rest of the paper is organized in the following way. In Section 2, we present the overview framework of rCom and key technical details of the system. In Section 3 we summarize the results of applying rCom to two different single cell transcriptomics data sets produced by the bone biology field. Section 4 is Discussion and Conclusion.

2 System and Methods

The overview of the rCom framework is given in Fig. 1. Fig 1a shows use of three different types of databases, TF-target database, ligand-receptor pairs database and gene signaling regulated route databases, which are used as prior knowledge in building ligand-receptor (denoted as L-R from here on) communication routes. These source databases are downloaded in text or KGML format and are parsed to build a network graph where ligands, receptors, genes, and TFs become nodes and their regulated relationships are formed into edges interconnecting nodes. In rCom’s graph structure, edges are directional encoding either activation or inhibition relationships (Fig.1a, 1c).

Fig 1b shows the use of three types of input data, single cell gene expression matrix which is obtained through quality control, normalization, scaling and log transform using SCANPY[4]. In the matrix, columns indicate cells and rows indicate genes. After log transformed, rCom treats gene with negative value as inhibited genes and gene with positive value as activated genes.



**Fig. 2 rICRN generated by rCOM using Acta2-lineage cell in osseointegration.**

Diagram, calendar

Description automatically generated

Fig 1. The overview architecture of rCom. a. The three categories of biological knowledge databases used in the rCom. b. The input data for rCom including: (i). preprocessed scRNA-seq with real cell group label for inferencing inter cell type communication. (ii). preprocessed scRNA-seq with randomized group label for permutation test. (iii). Cell group label meta data file. c. The matrix of communication score between cell type. d. The communication route pairs extract from 3 type databases. e. Route-wised communication scored in each cell. f. The post analysis of the cell types communication include: inferenced rICRN, summary table including cell type with permutation p-value, violin plot for the ligand-receptor expression value.

2.1 Cell-Cell communication routes

Fig 1d provides the detailed view of the conceptually depicted cell-cell communication. This model labels two interacting cells as “Secretor” denoting the cell secreting a ligand and “Receiver” denoting the cell in which the ligand binds to its cognate receptor. This binding is represented as the green dashed line in Fig 1d. We call the two cells as an L-R pair. We show how we generate communication routes from four different biological knowledge databases including: Harmonizome’s curation of TF targets[5] , CellChatDB [6] , CellTalkDB [7] and KEGG [8].

As illustrated in Fig 1d, each L-R pair is to combine two routes: (i) upstream ligand route (ULR) and (ii) downstream receptor route (DRR). Upstream ligand route captures the signals which are cell-secreted ligands produced when the transcription factor (TF) binds to the target genes and for this reason the cell that secrets ligands is labelled as Secretor. ULR usually starts from a TF and follows the path leading to the generation of ligands. Identifying ULR is done by first applying our route-finding program rPAC on KEGG pathways and then using Harmonizome’s TF targets. Identifying DRR is also done by applying rPAC to KEGG using the preprocessed gene expression matrix. Connecting ULR and DRR is done by exploring all possible combinations matching up putative ligand and receptor pairs from two ligand-receptors pair databases: CellChatDB and CellTalkDB. This step also uses rPAC scoring system that is readily available to us. When both CellChatDB and CellTalkDB are used, total 104,004 communication route pairs are examined which are statically stored as a graph structure in rCom.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

**Table 1. The significant communication routes pairs with score higher than threshold (8.5).**

2.2 Node expected value and evaluation value

*2.2.1 Node expected value.* A node expected value () is set to either +1 or -1 to indicate if the node is considered up-regulated or down-regulated in the route being examined. Each node in the route is assigned an expected value using a value propagation method which starts from ligands in ULR or receptors in DRR. The equation of this method is given in Eq 1.

(1)

In this equation, stands for the nodes from the starter node (ligand in ULR and receptor in DRR.) and stands for the edge expectation value between node and node. The value of is assigned using Eq 2.

*2.2.2 Node evaluation value.* Node evaluation value () is assigned based on the node expected value and gene expression value in cell . The evaluation value is computed using the Eq 3.

stands for the preprocessed expression value for gene in node of cell .

2.3 Communication route score

A communication route () score () denotes the probability that the cell communicates with other cells through the communication route . A cell with high ULR score means that the Secretor cell has a high probability to secret ligands which are also discovered highly regulated in the specific route j. A cell with a high DRR score means that the Receiver cell has a high probability to receive signals through its receptor and to further regulate its downstream genes including the TF included in DRR. The score is assigned using the equation given in Eq. 4a, 4b.

where is the number of nodes in the communication route , and and are the node evaluation value of ligand node or receptor node in ULR or DRR. Since we assume the cell-cell communication mostly relies on the ligand and receptor, a hyperparameter is introduced to adjust the effect of ligand or receptor. The choice of can be “context-dependent” ranging between 0 and 1 and it can be determined empirically like hyperparameters of machine learning models.

2.4 Inferring cell type communication

To infer the inter and intra-cellular communications, we introduce a heuristic rule which can model two distinct types of cell-cell communication known as autocrine signaling and paracrine signaling. In autocrine signaling, a cell secretes a messenger molecule (ligand) that binds to receptors on the same cell. In paracrine signaling, the ligand binds to receptors on a different cell [9]. A strong autocrine may lead to a weak paracrine and vice versa. This rule can be easily modified to account for other scenarios such as a cell is activated by both autocrine and paracrine. The communication strength score between the cell type and through the communication route pair is computed using Eq.5.

Here and represent the ULR and DRR score in cell group and are computed using Eq. 6a and 6b. and stand for the score of communication route for cell in cell group .

By checking the communication score , it is possible to rank the cell type communication strengths among all possible pairs of cell types in the data set.

2.5 Identification of statistically significant cell type communication routes

Next is to perform a permutation test to identify which cell type pair communication is more likely than others. The permutation test is done by randomly permuting the cell type labels 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. 7

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.

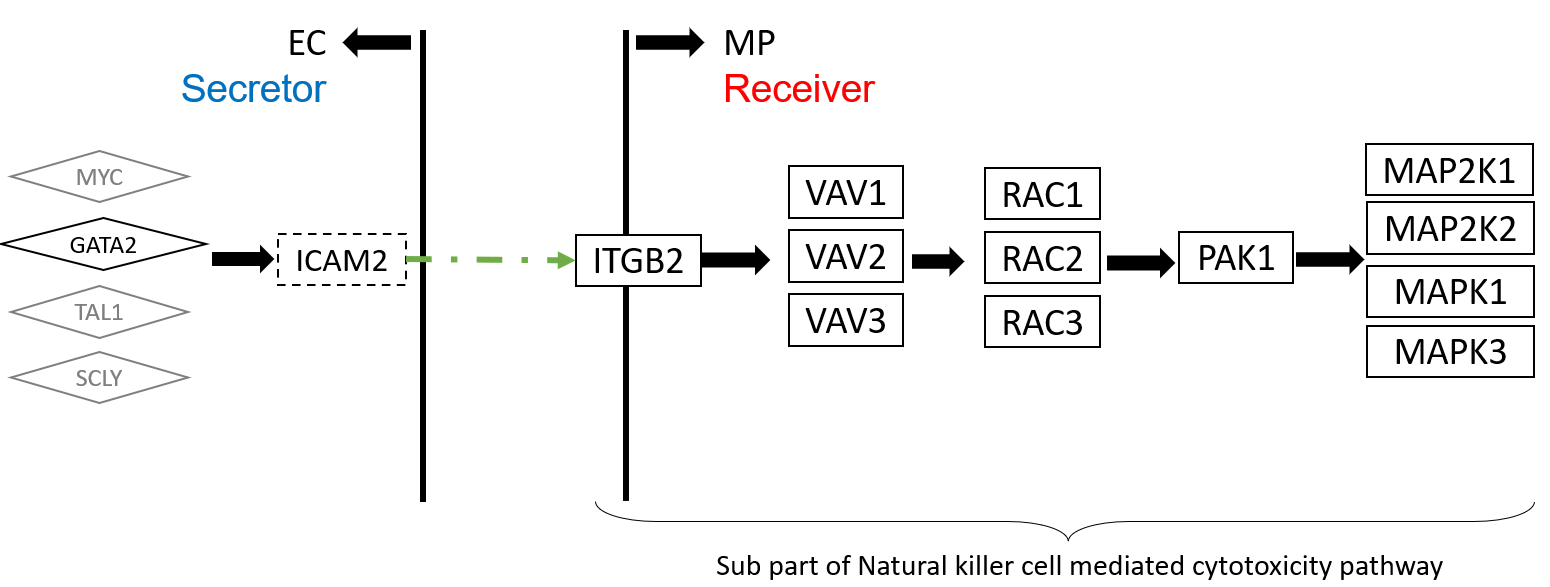
3 Experiment results

The single cell transcriptomics dataset we applied rCOM aims to analyze how bone marrow stromal cell subtype populations contribute to bone formation around metal implants [10]. After the QC preprocessing and data formatting, our method creates the scRNA-seq data matrix made up of 4397 cells with 31053 features (genes). Vesprey et al. provide the labels of each cell subgroup such as: a) cell of skeletal muscle (CSM), b) chondrocyte (CC), c) endothelia cell (EC), d) macrophage (MP), e) osteoblast (OB), f) smooth muscle cell (SMC) and g) stromal cell of bone marrow (SBM).



**Fig 4. The violin plot of ICAM2 and ITGB2 gene expression value in each cell type.**

The analysis outcome of applying rCOM to this dataset is summarized in Tab 1. It shows a small portion of discovered communication routes whose route score threshold > 8.5 for various of L-R pairs. Fig 2 shows the network rendering of the generated communication routes using Pyvis. Here each node denotes a cell type and the edge arrow signifies the signaling direction (e.g., from ligand in ULR to receptor in DRR). For example, Fig 2 includes two significant route pairs identified between EC and MP where EC is the secretor and the MP the receiver. The edge label is given to show the number of the significant communication pairs identified between the two involved cell subtypes. From the networks, we notice that the edges from SBM to EC, from OB to EC and from CC to EC have the greater number of significant communication routes. Among these, the communications between EC and MP seems well studied in the field [11]. In Fig 3, we elaborate the particular route (Id: 70072, as highlighted in Tab. 1) as rCom identified significant route between EC and MP. This figure shows the identification of the ULR interconnecting the transcription factor GATA2 and its target ICAM2 in EC. The route 70072 basically suggests that the secreted ICAM2 by EC may bind to its receptor ITGB2 in MP. ICAM2, a member of intercellular adhesion molecule family is generally known to bind to the leukocyte adhesion LFA-1 protein. But in this context of bone marrow stromal cell populations designed to form bone around metal implants, the discovery suggests that ICAM2 may more likely bind to ITGB2. In fact, the role of ITGB2 in differentiation of osteoblast precursor cells has been reported by Kim and Adchi 2019 [12]. Finally, the DRR discovered by rCom suggests that ITGB2 may regulate its downstream genes (e.g., VAV1, RAC1, etc.) and eventually activate the MAPK family genes as shown at the end of the route in Fig 3. The violin plots for gene expression of ICAM2 and ITGB2 in different cell groups are shown in Fig 4. ICAM2 is highly expressed in EC but scarcely expressed in MP while ITGB2 is highly expressed in MP but scarcely in EC, supporting the idea that the route 70072 signifies a paracrine signaling activity. Our literature survey also finds that in their 2011 work Zhang et al. [13] identified the communication between EC and MP through ICAM family gene, although not the details provided in the route 70072 were reported in their work.



**Fig 3. The upper stream and downstream genes in the communication route pair id:70072**

4. Discussion and conclusion

Developing rCom was motivated by inferring the cell type communication by utilizing signaling/molecular pathway routes identifiable when the appropriate ligand-receptor pairs from two interacting cells are mapped. Using the example from our case study, the communication between EC and MP have been reported in the context of muscle cell biology. Our analysis suggests that a similar regulatory relationship may occur between SMC and PL in bone making. Like in many computational methods, such finding should be considered as a potential lead, and when such pattern reoccurs in high frequency in different, related analyses, the lead should merit experimental validation.

For future work, we believe rCom is well poised to handle spatial genomics data (e.g., seqFISH/seqFISH+, MERFISH, Visium, etc.) in which cells’ proximity information can be combined with their transcriptomic/protein data to improve the accuracy of cell-cell communication and their regulatory relationships at single cell resolution.

ACKNOWLEDGMENTS

Research reported in this work was supported in part by NIH/NICHD Grant No.1R01HD098636-01. Its contents are solely the responsibility of the authors and donot necessarily represent the official views of the NIH.

**REFERENCES**

[1] D. A. Skelly et al., “Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart,” Cell Reports, 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 Reports, vol. 25, no. 6, pp. 1458-1468.e4, Nov. 2018,

[3] Y. Hu, T. Peng, L. Gao, and K. Tan, “CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data,” Science Advances, vol. 7, no. 16, Apr. 2021,

[4] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: Large-scale single-cell gene expression data analysis,” Genome Biology, vol. 19, no. 1, pp. 1–5, Feb. 2018, doi: 10.1186/S13059-017-1382-0/FIGURES/1.

[5] A. D. Rouillard et al., “The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins,” Database (Oxford), vol. 2016, 2016, doi: 10.1093/DATABASE/BAW100.

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

[7] 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,” Briefings in Bioinformatics, vol. 22, no. 4, Jul. 2021,

[8] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” 2000.

[9] A. Caicedo, “PARACRINE AND AUTOCRINE INTERACTIONS IN THE HUMAN ISLET: MORE THAN MEETS THE EYE,” Semin Cell Dev Biol, vol. 24, no. 1, p. 11, 2013,

[10] 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.

[11] H. He et al., “Endothelial cells provide an instructive niche for the differentiation and functional polarization of M2-like macrophages,” Blood, vol. 120, no. 15, p. 3152, Oct. 2012

[12] J. Kim and T. Adachi, “Cell Condensation Triggers the Differentiation of Osteoblast Precursor Cells to Osteocyte-Like Cells,” Frontiers in Bioengineering and Biotechnology, vol. 7, p. 288, Oct. 2019,

[13] 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.