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

Pujan Joshi
Computer Sci. and Engr. Dept.,
Univ. of Connecticut, Storrs, USA
email@email.com

Chenyu Zhang
Computer Sci. and Engr. Dept.,
Univ. of Connecticut, Storrs, USA
chenyu.zhang@uconn.edu

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

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

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

ABSTRACT

The mapping of ligand-receptor pairs is the cornerstone of understanding complicated intercellular interactions. 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. The rCom framework examines combinatoric ways of integrating the partially known relationships against the cohorts of gene expression datasets obtainable through subtyped cells. 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 two case studies 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. Zhou et al. [3] suggest identifying ligand-receptor pair signal between cell types is needed in addition to genetic mutations or intracellular signaling to better understand cancer progression. Vento-Tormo et al. [4] produce a repository of ligand-receptor complexes which could potentially explain interaction between decidua and placenta during the early human pregnancy. 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. Wang et al. [5] introduce SoptSC which can predict cell-cell communication networks by constructing a structured cell-to-cell similarity matrix and. Browaeys et al. [6] report NicheNet which aims to identify not only ligand-receptor pairs of interacting cells but also the genes downstream of the identified pairs. Hu et al [7] 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.

Among these more recent cell-cell communication network discovery methods based on single cell transcriptomics data, our system, namely, rCom, is similar to NicheNet in the sense that both uses prior knowledge in identifying cell-cell communication network. One key difference between the two systems is in the way the prior knowledge curated into signaling/molecular pathways is used for the analysis. 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. In contrast, NicheNet does not

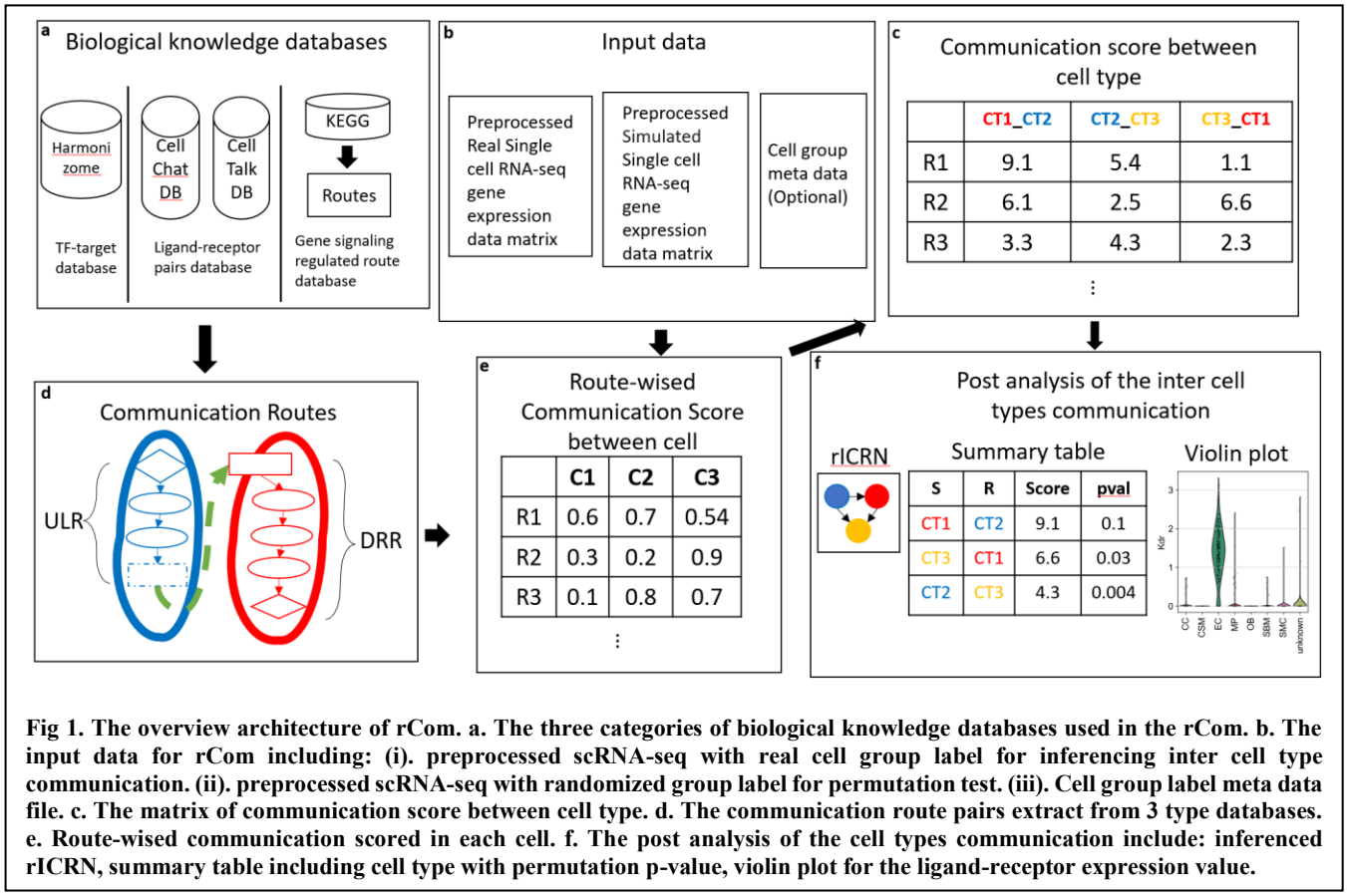

Fig 1. The overview architecture of rCom. a. The three categories of biological knowledge databases used in the rCom. b. The input data for rCom including: (i). preprocessed scRNA-seq with real cell group label for inferencing inter cell type communication. (ii). preprocessed scRNA-seq with randomized 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.

use gene regulation notions such as activation/inhibition in pathway scoring and simply uses aggregation of pathways sources based on gene expression values.

Creating rCom is based on our earlier works using "routes" of pathways, i.e., branches of curated pathways (e.g., KEGG) as the unit for the analysis as opposed to the entire assembled genes of a pathway. This idea is based on the observation that the relationships between signaling/molecular pathways and biological processes are N:M, and thus the granularity of pathway analysis should be "routes", i.e., which particular branch of a pathway could be responsible for exerting the biological process under question. Our previous route-based pathway analysis works include BioTarget which uses the route-based pathway method to extend existing Th1/Th2 pathways using TCGA data sets [8] and rPAC [9] which uses a transcription factor (TF) centric analysis to identify more refined cancer subtype signatures [9]. Development of rCom was motivated to use this route-based pathway analysis to estimate how two cells may interact with each other through ligand-receptor pairing.

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[10]. 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.

2.1 Cell-Cell communication routes

Fig 2 provides the detailed view of the conceptually depicted cell-cell communication given in Fig 1d. 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 2. We call the two cells as an L-R pair. We show how we generate communication routes from four different biological

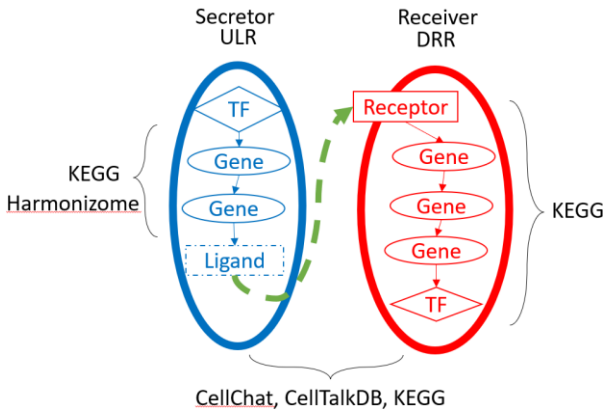

Fig 2. The overview of the communication route. The blue route is ULR, and the red route is DRR. The ULR is generated from rPAC routes and Chip-seq database. The DRR is generated from rPAC routes. The ligand and receptor pairs are derived from CellChatDB, CellTalkDB and KEGG.

knowledge databases including: Harmonizome's curation of TF targets[11] , CellChatDB [12] , CellTalkDB [13] and KEGG [14].

As illustrated in Fig 2, 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.

2.2 Node expected value and evaluation value

2.2.1 Node expected value. A node expected value (E_k) 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.

$$E_k = \begin{cases} 1 & k = 1 \\ E_{k-1} * e_k & other\ wise \end{cases} \quad (1)$$

In this equation, k stands for the k_{th} nodes from the starter node (ligand in ULR and receptor in DRR.) and e_k stands for the edge expectation value between k_{th} node and $k_{th} - 1$ node. The value of e_k is assigned using Eq 2.

$$e_k = \begin{cases} +1 & the\ edge\ is\ activation \\ -1 & the\ edge\ is\ inhibition \end{cases} \quad (2)$$

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

$$V_{k_i} = \begin{cases} abs(EV_{ki}) & EV_{ki} * E_k \geq 0 \\ 0 & EV_{ki} * E_k < 0 \end{cases} \quad (3)$$

EV_{ki} stands for the preprocessed expression value for gene in node k of cell i.

2.3 Communication route score

A communication route (ULR_j and DRR_j) score (SL_{ij} and SR_{ij}) denotes the probability that the cell i communicates with other cells through the communication route j. 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.

$$SL_{ij} = \begin{cases} w * VL + (1-w)\frac{1}{n_i - 1}\sum_k^{n_i - 1} V_k + 1 & if\ VL > 0 \\ 1 & if\ VL \leq 0 \end{cases} \quad (4a)$$

$$SR_{ij} = \begin{cases} w * VR + (1-w)\frac{1}{n_i - 1}\sum_k^{n_i - 1} V_k + 1 & if\ VR > 0 \\ 1 & if\ VR \leq 0 \end{cases} \quad (4b)$$

where n_i is the number of nodes in the communication route i, and VL and VR 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 w is introduced to adjust the effect of ligand or receptor. The choice of w 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 [15]. We control the analysis between autocrine and paracrine by introducing a mass function. 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 α and β through the communication route pair j is computed using Eq.5.

$$C_{j\alpha\beta} = \frac{SL_{\alpha j} * SR_{\beta j}}{SL_{\alpha j} * SR_{\alpha j} + SL_{\beta j} * SR_{\beta j}} \quad (5)$$

Here $SL_{\alpha j}$ and $SR_{\alpha j}$ represent the ULR and DRR score in cell group α and are computed using Eq. 6a and 6b. SL_{kj} and SR_{kj} stand for the score of communication route j for cell k in cell group α.

$$SL_{\alpha j} = \sqrt[n]{SL_{1j} * \ldots * SL_{kj} * \ldots * SL_{nj}} \quad (6a)$$

$$SR_{\alpha j} = \sqrt[n]{SR_{1j} * \ldots * SR_{kj} * \ldots * SR_{nj}} \quad (6b)$$

By checking the communication score ($C_{j\alpha\beta}$), it is possible to rank the cell type communication strengths among all possible pairs of cell types in the data set. The higher score represents a higher probability the communication may take place between the cells in two cell types through the communication route under consideration.

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 ($C_{j\alpha\beta}{}'$) between cell type α and β through communication route pair j. The p-value of each communication score ($C_{j\alpha\beta}$) is calculated using Eq. 7

$$Pvalue = 1 - \frac{\{\#m | C_{j\alpha\beta}^{(m)\prime} \leq C_{j\alpha\beta}, m = 1,2,\ldots,M\}}{M} \quad (7)$$

where the score $C_{j\alpha\beta}^{(m)\prime}$ 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

We experimented rCom with two sets of independent single cell transcriptomics data related to bone biology, one from the Single Cell Portal managed by The Broad Institute of MIT and Harvard, and one from NCBI GEO. Two sets of independent but related data sets are used to demonstrate the generality of rCOM and to see if derivation of cell-cell communications can be repeated in a similar manner in different single cell transcriptomics datasets.

3.1 Experiments with Acta2-lineage cells in osseointegration dataset

The first single cell transcriptomics dataset we applied rCOM aims to analyze how bone marrow stromal cell subtype populations contribute to bone formation around metal implants [16]. 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. [16] provide the labels of each cell subgroup such as: a) cell of skeletal muscle (CSM) (269 cells), b) chondrocyte (CC) (545 cells), c) endothelia cell (EC)

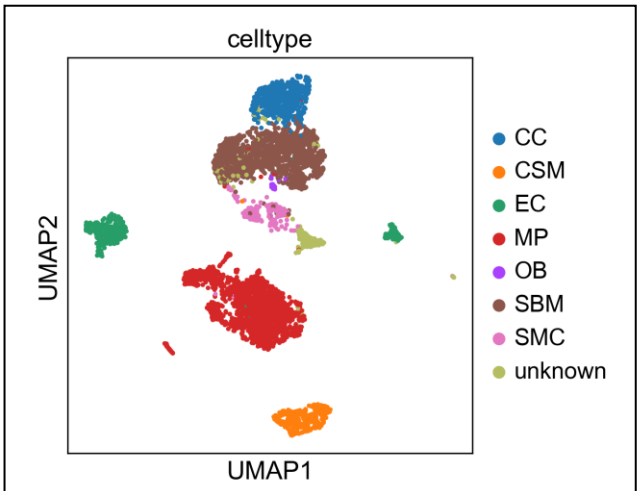

Fig 3. The reproduce 2-D visualization of UMAP, colored by cell label provided by studies.

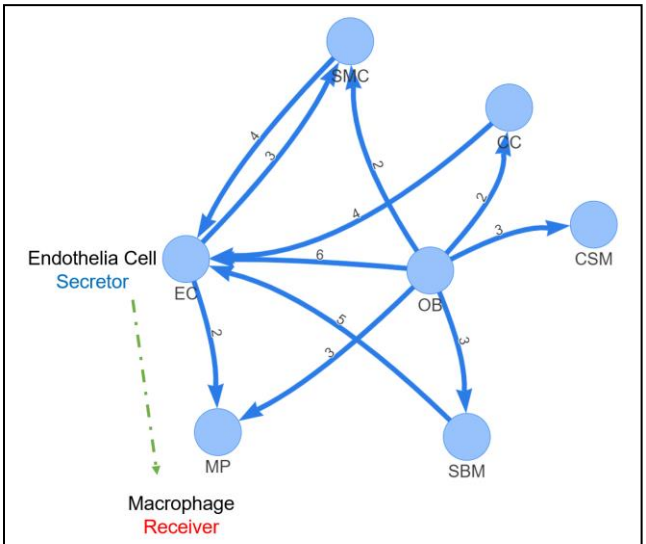

Fig. 4 rICRN generated by rCOM using Acta2-lineage cell in osseointegration. The label of edge shows the number of communication routes between two cell groups

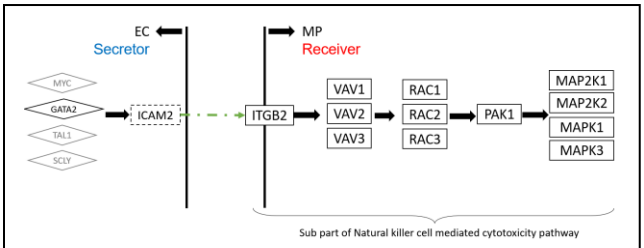

Fig 5. The upper stream and downstream genes in the communication route pair id:70072

(394 cells), d) macrophage (MP) (1608 cells), e) osteoblast (OB) (19 cells), f) smooth muscle cell (SMC) (191 cells) and g) stromal cell of bone marrow (SBM) (1371 cells). We were able to reproduce this particular subgrouping of the study cells using the 2-D UMAP visualization as shown in Fig 3, similar to what has been reported in [16]. This exercise provide confidence in our use of reported cell type labels.

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 4 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 4 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 [17]. In Fig 5, 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 particular 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 [18]. 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 5. The violin plots for gene expression of ICAM2 and ITGB2 in different cell groups are shown in Fig 6. 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. [19] 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. Lastly, we point out that there are multiple other transcription factors known to regulate ICAM2 such as MYC, TAL1 and SCLY besides GATA2 as shown in Fig 5. The ULR picked by rCom includes only GATA2 and this fact can also be verified from the violin plots of gene expression for the upper stream TFs of ICAM2 as shown in Fig 7. GATA2's expression level stands out in EC.

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

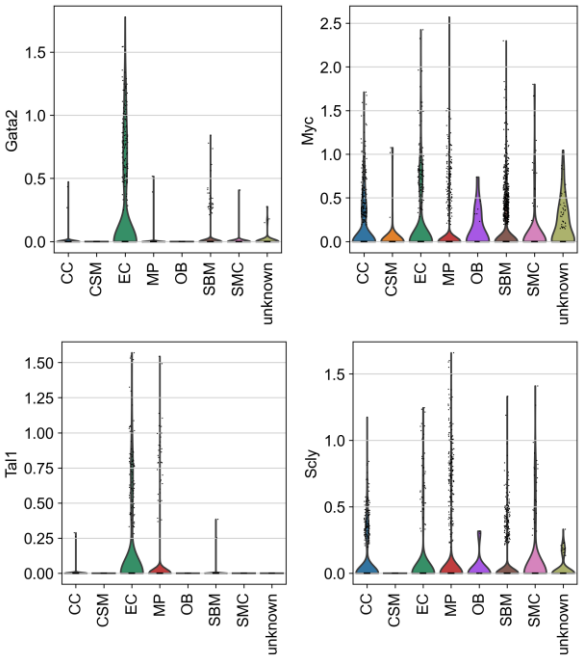

Fig 7. The violin plot of COL18A1 and ITGB3 gene expression value in each cell type.

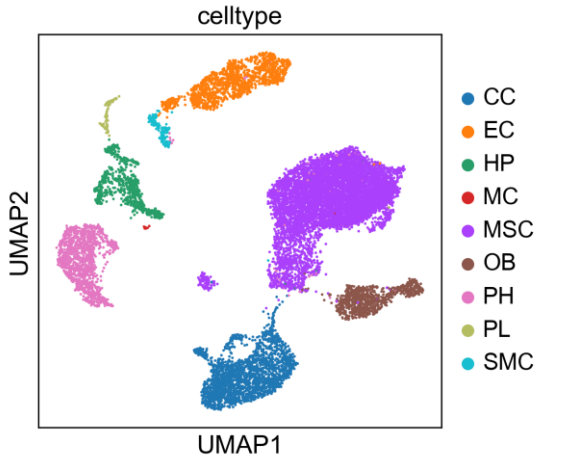

Fig 8. The reproduce 2-D visualization of UMAP, colored by cell label provided by the study.

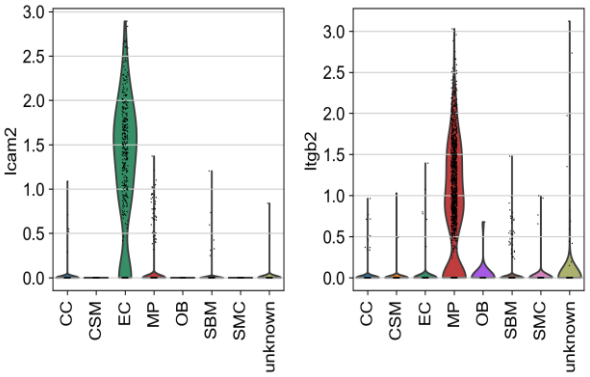

Fig 6. The violin plot of ICAM2 and ITGB2 gene expression value in each cell type.

The second single cell transcriptomics dataset is from a study [20] which explores osteogenesis in adult mice and shows that the regulatory mechanism of bone making may rely on Notch signaling. This study data is obtained from bone stromal cells (GEO GSE152285). After preprocessing, the scRNA-seq data matrix has 17015 cells and 15046 features (genes). The authors of the study also provide the labels of each cell subtype such as CC (2270 cells), EC (1677 cells), Hematopoietic (HP) (849 cells), mesenchymal stromal cells (MSC) (9168 cells), osteoblast (OB) (971 cells), platelets (PL) (232 cells), proliferating Hem (PH) (1532 cells), smooth muscle cells (SMC) (238 cells), MC (78 cells). We repeated their cell subtyping and show in Fig 9 that our UMAP analysis approximately reproduces the authors' 2-D cell subtype visualization.

The outcome of applying rCom to this dataset is summarized in Tab 2 in which only a portion of communication routes with their route score > 9 is shown. The network rendering of the generated communication routes using Pyvis is given in Fig 9. Noticeable in the figure is that the communication between SMC and PL has the largest number (11) of significant route pairs whereas the communication from MC to PH has only 2 route pairs. In fact, the interaction between SMC and PL appears to be well documented

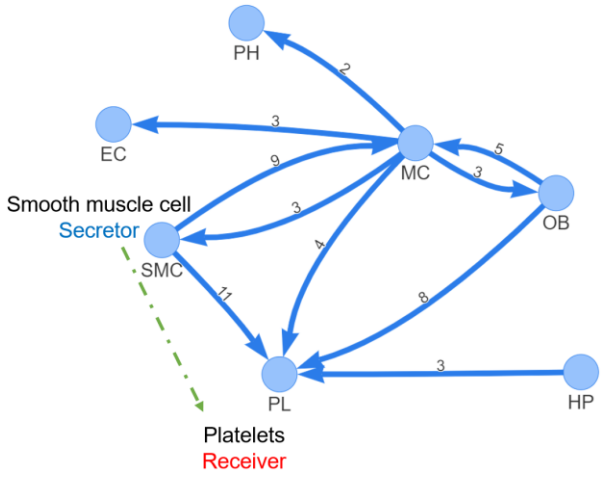

Fig. 9 rICRN generated by rCOM using bone marrow stromal dataset scRNA-seq of 11 weeks old mice dataset

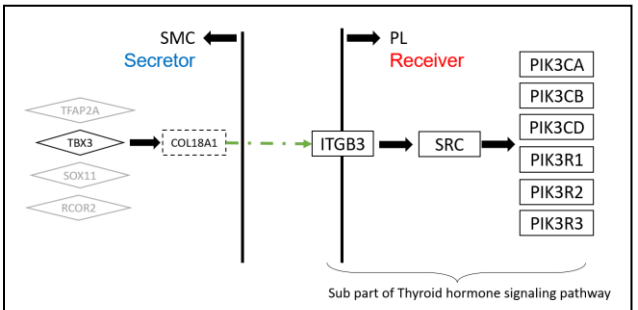

Fig 10. The upper stream and downstream genes in the communication route pair id 95696

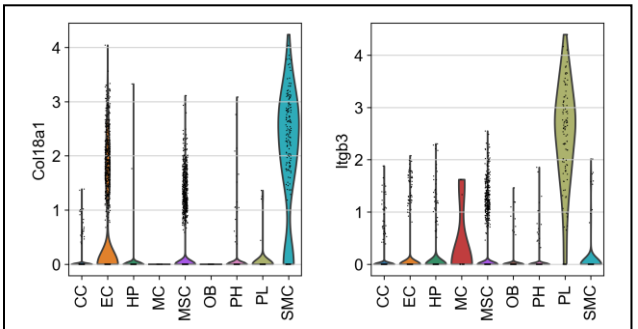

Fig 11. The complete view of communication route pairs

in the bone biology field [21]. We elaborate on this specific route (ID: 95696) which acquired the highest route score 12.64 by rCom as shown in Tab 2. The pictorial presentation of this route is shown in Fig 10 which highlights SMC as the Secretor and PL as the Receiver. In its ULR identification, the transcription factor TBX3 is estimated to target COL18A in SMC, and then this COL18A1 binds to ITGB3 in PL. In the DRR identification,

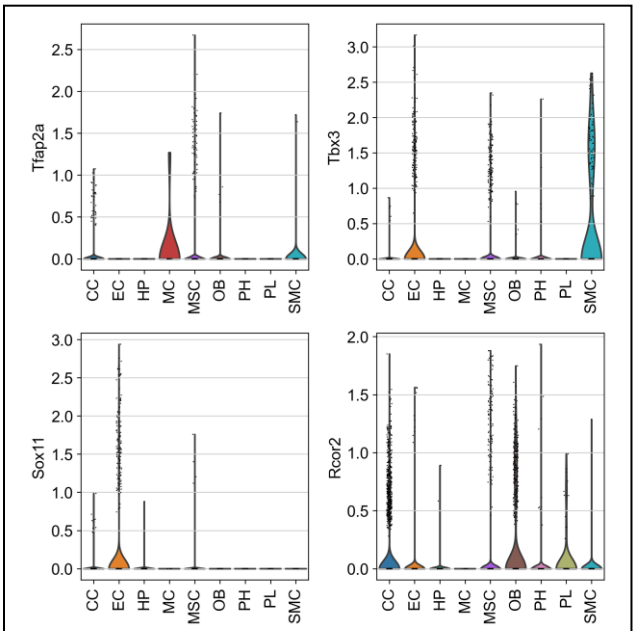

Fig 12. The violin plot for upper stream gene expression value of COL18A1 and ITGB3.

S	R	R id	Score	ULR	Ligand	Receptor	DRR	Pval
OB	EC	82280	9.83	RUNX2	IBSP	ITGB3	PLCD3, PLCB1, PLCE1, PLCB2, PLCB3, PLCB4, PLCD1, PLCG1, PLCG2, PLCD4, PLCZ1, C00165, PRKCA, PRKCB, PRKCG, MAPK1, MAPK3, STAT1	<0.0005
OB	MP	96177	9.53	RUNX2	IBSP	ITGAV	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
OB	SMC	96177	9.18	RUNX2	IBSP	ITGAV	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
OB	SBM	96177	9.05	RUNX2	IBSP	ITGAV	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
EC	***MP***	***70072***	***9.04***	***GATA2***	***ICAM2***	***ITGB2***	***ITGAL, VAV3, VAV1, VAV2, RAC1, RAC2, RAC3, PAK1, MAP2K1, MAP2K2, MAPK1, MAPK3***	**<0.0005**
OB	EC	6452	8.68	SMAD3	TNC	PTPRB	CTNNB1, CTNNA1, CTNNA2, CTNNA3, ACTB, ACTG1	<0.0005
EC	OB	77882	8.66	CREM	GNAI2	LPAR3	GNAS, ADCY1, ADCY2, ADCY3, ADCY5, ADCY6, ADCY7, ADCY8, ADCY9, ADCY4, C00575, RAPGEF3, RAPGEF4, RAP1A, RAP1B	<0.0005
OB	EC	103917	8.65	RUNX2	VEGFA	KDR	PLCG1, PLCG2, C00076, PRKCA, PRKCB, PRKCG, SPHK2, SPHK1, HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005
OB	EC	4562	8.64	RUNX2	SEMA5A	MET	MAPK1, MAPK3, SNAI2, SNAI1	<0.0005
OB	EC	103718	8.63	CTNNB1	PDGFC	KDR	PLCG1, PLCG2, C00076, PRKCA, PRKCB, PRKCG, SPHK2, SPHK1, HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005
OB	EC	89677	8.55	RUNX2	IBSP	ITGAV	HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3, STAT1	<0.0005
SMC	EC	103686	8.53	PBX1	VTN	KDR	PLCG1, PLCG2, C00076, PRKCA, PRKCB, PRKCG, SPHK2, SPHK1, HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005

Table 1. The significant communication routes pairs with score higher than threshold (8.5). S stands for secretor and R stands for receiver.

ITGB3 regulates its downstream genes and activates the PIK family genes in PL as shown at the end of the route given in Fig 10. The gene expression value distribution for COL18A1 and ITGB3 is shown for all identified cell subtypes using violin plots in Fig. 11. Noticeable here is that among all cell subtypes, COL18A1 and ITGB3 are predominantly expressed in the SMC and ITGB3, respectively. Our literature survey reveals that according to Misra et al. [22] SMC and PL are interacting with each other and ITGB3 plays an important role in smoothing muscle-derived atherosclerotic plaque cells. A similar report about the interface between SMC and PL is also given by Inoue et al 2015 [23] who state that smooth muscle cells stimulate platelets in atherothrombosis. We show the gene expression value distribution for all the TFs that are known to regulate the COL18A1 including TFAP2A, TBX3, NUCKS1, SOX11 and RCOR2 in Fig 12. Only TBX3 are highly expressed in the SMC suggesting a support for rCOM's pick of TBX3 as the upstream for COL18A. Lastly, Sipola et al (2009) [24] report that COL18A1 affects osteoblast behavior. It remains to be seen how and if COL18A1 may play a role in interfacing SMC and PL in osteoblast driven bone making.

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. Our system can be compared with CtyoTalk which is one of recently reported cell-cell communication discovery systems. Like in rCom, CtyoTalk also produces a signal transduction network between cell types. But one key difference between CytoTalk and rCom is if any prior knowledge is used or not for the inference. The network inferred by CytoTalk is strictly data driven—a form of unbiased analysis. In contrast, rCom aims to extend the "existing" curated network by introducing de novo pathway routes as best guesses. The assumption behind this use of prior knowledge is that many curated resources record regulatory relationships either computationally predicted or experimentally derived. The chance these relationships repeat in related biological context should be higher than random, and the single cell gene expression data set used for the analysis can be seen as evidences supporting the "context-specific" extension of the derived cell-cell communication. Using the example from our second case study, the communication between SMC and PL 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.

The current version of with rCom was developed using Harmonizome's curation of TF targets based on ENCODE ChIP-Seq, CellTalkDB and CellChatDB for curated ligand-receptor pairs, and KEGG for curated signaling pathways. However, it can be easily extended to use other databases such as BioGRID [25] and STRING [26]. 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

S	R	R id	Score	ULR	Ligand	Receptor	DRR	Pval
SMC	***PL***	***95696***	***12.64***	***TBX3***	***COL18A1***	***ITGB3***	***SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3***	***<0.0005***
SMC	MC	80134	11.58	NFIB	LGALS1	PTPRC	LCK, ZAP70, LAT, GRB2, SOS1, SOS2, RASGRP1, HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3, FOS, JUN	<0.0005
SMC	PL	95739	11.03	PPARG	COL4A2	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
SMC	PL	95421	10.34	NR3C1	MFGE8	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
SMC	PL	95367	9.7	PPARG	NID1	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
MC	SMC	72456	9.63	RELA, RELB	IL1B	IL1R1	IRAK1, IRAK4, MYD88, TRAF6, MAP3K7, TAB1, TAB2, TAB3, IKBKG, CHUK, IKBKB, NFKBIA, NFKB1, RELA	<0.0005
OB	PL	82280	9.6	RUNX2	IBSP	ITGB3	PLCD3, PLCB1, PLCE1, PLCB2, PLCB3, PLCB4, PLCD1, PLCG1, PLCG2, PLCD4, PLCZ1, C00165, PRKCA, PRKCB, PRKCG, MAPK1, MAPK3, STAT1	<0.0005
OB	PL	82619	9.45	KLF4	FN1	ITGB3	PLCD3, PLCB1, PLCE1, PLCB2, PLCB3, PLCB4, PLCD1, PLCG1, PLCG2, PLCD4, PLCZ1, C00165, PRKCA, PRKCB, PRKCG, MAPK1, MAPK3, STAT1	<0.0005
MC	EC	72456	9.44	RELA, RELB	IL1B	IL1R1	IRAK1, IRAK4, MYD88, TRAF6, MAP3K7, TAB1, TAB2, TAB3, IKBKG, CHUK, IKBKB, NFKBIA, NFKB1, RELA	<0.0005
CC	PL	90899	9.42	STAT3	TGM2	ITGB3	HRAS, KRAS, NRAS, RAF1, MAP2K1, MAP2K2, MAPK1, MAPK3, TP53	<0.0005
OB	PL	82963	9.41	CCN1	CCN1	ITGB3	PLCD3, PLCB1, PLCE1, PLCB2, PLCB3, PLCB4, PLCD1, PLCG1, PLCG2, PLCD4, PLCZ1, C00165, PRKCA, PRKCB, PRKCG, MAPK1, MAPK3, STAT1	<0.0005
SMC	PL	95503	9.39	TCF4	FBN1	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
OB	MC	69430	9.36	CCN1	CCN1	ITGB2	PTK2B, RAC1, RAC2, RAC3, PAK1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005
SMC	PL	95268	9.35	CCND1	VTN	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
HP	PL	95560	9.34	MYB	HSP	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
SMC	MC	69220	9.3	MEF2A	JAM3	ITGB2	PTK2B, RAC1, RAC2, RAC3, PAK1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005
MC	CC	72456	9.11	RELA, RELB	IL1B	IL1R1	IRAK1, IRAK4, MYD88, TRAF6, MAP3K7, TAB1, TAB2, TAB3, IKBKG, CHUK, IKBKB, NFKBIA, NFKB1	<0.0005
SMC	PL	95290	9.08	EP300	TGFB3	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
SMC	PL	96031	9.08	TRIM28	ITGAV	ITGB3	SRC, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3	<0.0005
MC	OB	72456	9.04	RELA, RELB	IL1B	IL1R1	IRAK1, IRAK4, MYD88, TRAF6, MAP3K7, TAB1, TAB2, TAB3, IKBKG, CHUK, IKBKB, NFKBIA, NFKB1	<0.0005
OB	PL	82885	9.04	KLF4	SPP1	ITGB3	PLCD3, PLCB1, PLCE1, PLCB2, PLCB3, PLCB4, PLCD1, PLCG1, PLCG2, PLCD4, PLCZ1, C00165, PRKCA, PRKCB, PRKCG, MAPK1, MAPK3, STAT1	<0.0005
SMC	MC	69430	9	CCN1	CCN1	ITGB2	PTK2B, RAC1, RAC2, RAC3, PAK1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005
SMC	PL	95963	9	CCN1	CCN1	ITGB3	PTK2B, RAC1, RAC2, RAC3, PAK1, MAP2K1, MAP2K2, MAPK1, MAPK3	<0.0005

Table 2. The significant communication routes pairs with score higher than threshold (9). S stands for secretor and R stands for receiver.

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 do not 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, 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," Scientific Reports 2017 7:1, 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," Nature 2018 563:7731, vol. 563, no. 7731, pp. 347–353, Nov. 2018, doi: 10.1038/s41586-018-0698-6.

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

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

[7] 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, doi: 10.1126/SCIADV.ABF1356/SUPPL_FILE/ABF1356_TABLE_S6.XLSX.

[8] T. H. Hoang et al., "BioTarget: A Computational Framework Identifying Cancer Type Specific Transcriptional Targets of Immune Response Pathways," Scientific Reports, vol. 9, no. 1, p. 9029, 2019, doi: 10.1038/s41598-019-45304-x.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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," Briefings in Bioinformatics, vol. 22, no. 4, Jul. 2021, doi: 10.1093/BIB/BBAA269.

[14] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," 2000.

[15] 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, doi: 10.1016/J.SEMCDB.2012.09.007.

[16] 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.

[17] 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, doi: 10.1182/BLOOD-2012-04-422758.

[18] 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, doi: 10.3389/FBIOE.2019.00288/BIBTEX.

[19] 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.

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

[21] R. Ross and L. Harker, "Platelets, endothelium, and smooth muscle cells in atherosclerosis," Adv Exp Med Biol, vol. 102, pp. 135–141, 1978, doi: 10.1007/978-1-4757-1217-9_8.

[22] 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.

[23] O. Inoue et al., "Vascular Smooth Muscle Cells Stimulate Platelets and Facilitate Thrombus Formation through Platelet CLEC-2: Implications in Atherothrombosis," PLOS ONE, vol. 10, no. 9, p. e0139357, Sep. 2015, doi: 10.1371/JOURNAL.PONE.0139357.

[24] A. Sipola, L. Seppinen, T. Pihlajaniemi, and J. Tuukkanen, "Endostatin affects osteoblast behavior in vitro, but collagen XVIII/endostatin is not essential for skeletal development in vivo," Calcif Tissue Int, vol. 85, no. 5, pp. 412–420, Nov. 2009, doi: 10.1007/S00223-009-9287-X.

[25] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets", doi: 10.1093/nar/gkj109.

[26] D. Szklarczyk et al., "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," Nucleic Acids Research, vol. 47, pp. 607–613, 2018, doi: 10.1093/nar/gky1131.