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
pujan.joshi@uconn.edu

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

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. 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. Hu et al [3] present that CytoTalk can generate a signal transduction network between a pair of cell types.

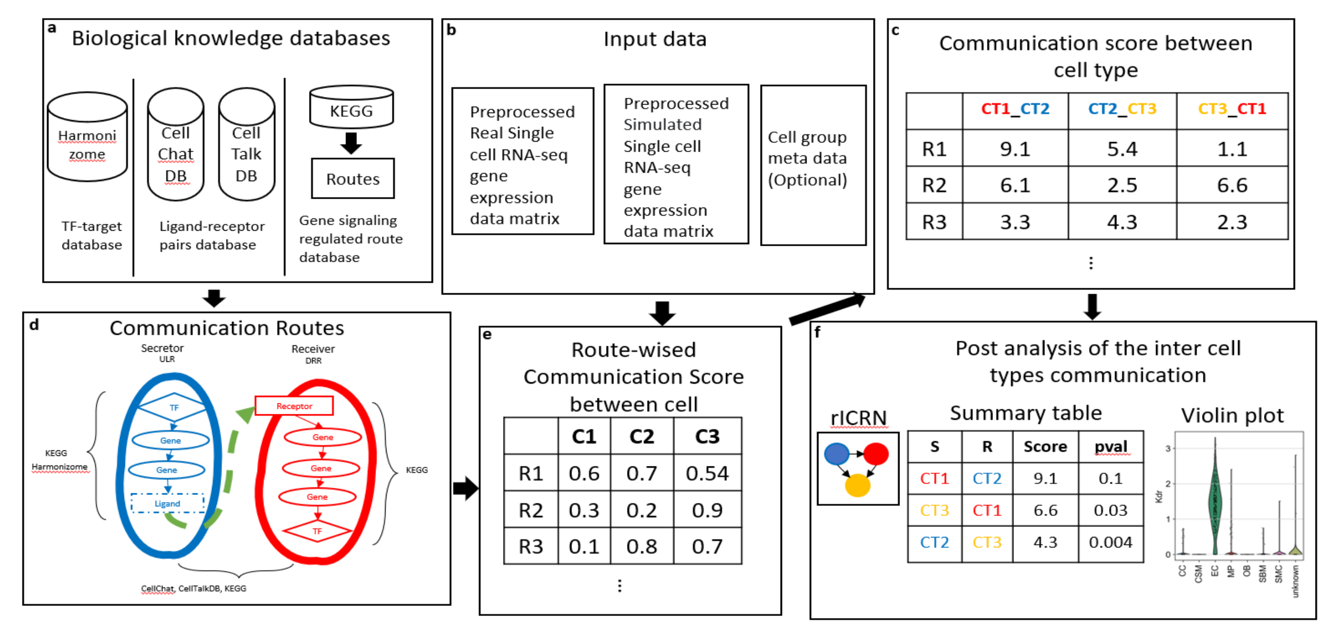
In rCom, inferring cell-cell interactions is decomposed into three parts, analysis of the upstream of the ligand, analysis of the downstream of receptors, and analysis of the most plausible ligand-receptor pairing. Our approach assumes subtyping cells are performed a priori and the labels for each cell 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. 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 [4] and rPAC [5] which uses a transcription factor (TF) centric analysis to identify more refined cancer subtype signatures.

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 (L-R) communication routes. These source databases 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. 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[6]. 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. 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 show how we generate communication routes from four different


a
Biological knowledge databases
Harmonizome
Cell Chat DB
Cell Talk DB
KEGG
Routes
TF-target database
Ligand-receptor pairs database
Gene signaling regulated route database
b
Input data
Preprocessed Real Single cell RNA-seq gene expression data matrix
Preprocessed Simulated Single cell RNA-seq gene expression data matrix
Cell group meta data (Optional)
c
Communication score between cell type
⋮
d
Communication Routes
Secretor
ULR
Receiver
DRR
KEGG
Harmonizome
TF
Gene
Gene
Ligand
Receptor
Gene
Gene
Gene
TF
KEGG
CellChat, CellTalkDB, KEGG
e
Route-wised Communication Score between cell
⋮
f
Post analysis of the inter cell types communication
rICRN
Summary table
Violin plot
Kdr
0
1
2
3
CC
CSM
EC
MP
OB
SBM
SMC
unknown

	CT1_CT2	CT2_CT3	CT3_CT1
R1	9.1	5.4	1.1
R2	6.1	2.5	6.6
R3	3.3	4.3	2.3

	C1	C2	C3
R1	0.6	0.7	0.54
R2	0.3	0.2	0.9
R3	0.1	0.8	0.7

S	R	Score	pval
CT1	CT2	9.1	0.1
CT3	CT1	6.6	0.03
CT2	CT3	4.3	0.004

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.

biological knowledge databases including: Harmonizome's curation of TF targets[7] , CellChatDB [8] , CellTalkDB [9] and KEGG [10]. 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). ULR 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 are 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.

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. 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.

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

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

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

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

2.3 Communication route score

A communication route ($\mathrm{ULR_j}$ and $\mathrm{DRR_j}$) score ($\mathrm{SL_{ij}}$ and $\mathrm{SR_{ij}}$) 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.

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

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

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 [11]. 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.

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

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

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.

2.5 Identification of statistically significant routes

Next step 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

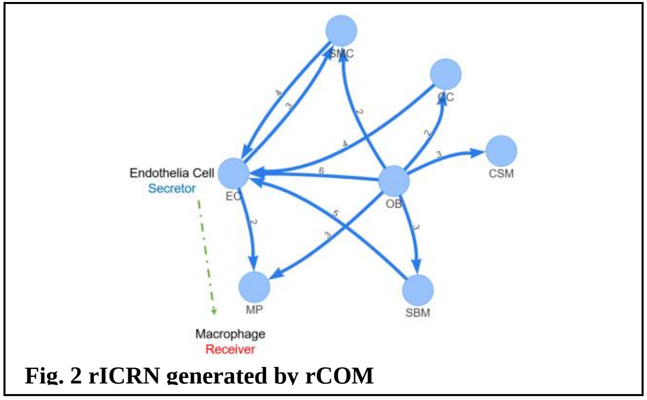


Fig. 2 rICRN generated by rCOM

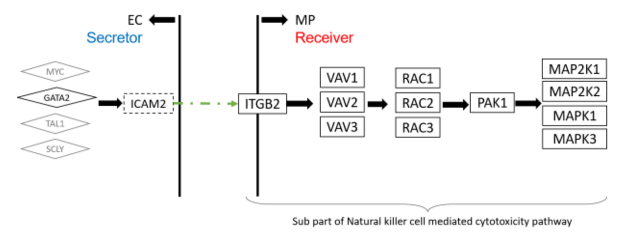


Fig 3. The upper stream and downstream genes in the communication route pair id:70072

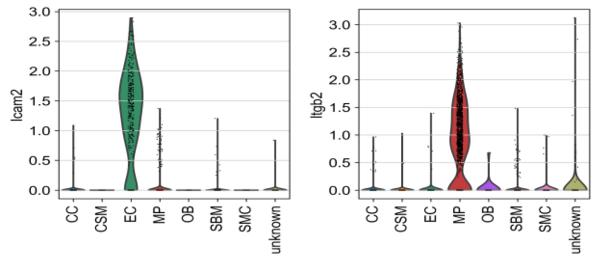


Fig 4. The violin plot of ICAM2 and ITGB2 gene expression value in each cell type.

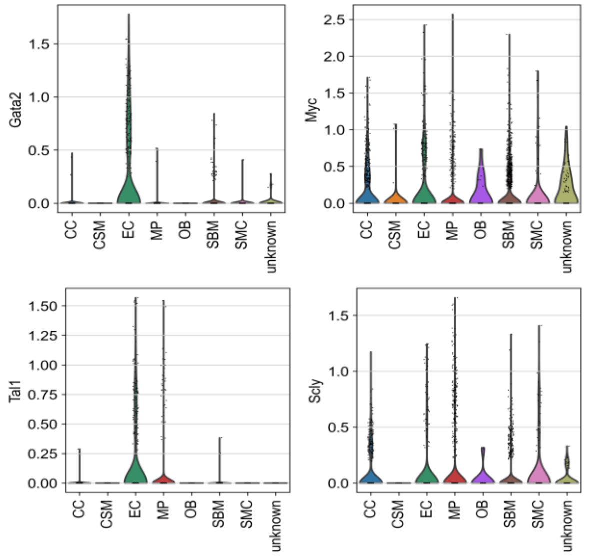


Fig 5. The violin plots of gene expression in route id: 70072.

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

where the score $C_{j\alpha\beta}^{(m)}{'}$ is the communication probability for the m-th permutation. M is the total number permutations (default is 100). The significant communications have p-value < 0.05.

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 [12]. 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).

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. Here each node denotes a cell type and the edge arrow signifies the signaling direction. 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 [13]. In Fig 3, we elaborate the 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 [14]. 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. [15] 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 3. 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 5. GATA2's expression level stands out in EC.

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. 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 EC and MP 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.

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

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 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] 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] 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,

[5] 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.

[6] 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.

[7] 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.

[8] 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,

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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,

[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,