LR hunting: A Random Forest Based Cell-cell Interaction Discovery Method for Single-cell Gene Expression Data

Min Lu[1], Yifan Sha[2], Tiago C. Silva[1], Antonio Colaprico[1], Xiaodian Sun[2], Yuguang Ban[1,2], Lily Wang[1,2,3,4], Brian D. Lehmann[5,6] and X. Steven Chen[1,2]*

[1]Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL, USA
[2]Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, FL, USA
[3]Dr. John T Macdonald Foundation Department of Human Genetics, Miller School of Medicine, University of Miami, Miami, FL, USA
[4]John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL, USA
[5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[6]Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA

*** Correspondence:**
X. Steven Chen
steven.chen@med.miami.edu

Keywords: random forests, ligand-receptor interaction, cell-cell interaction, cell-cell communications, single-cell RNA-seq

Abstract

Cell-cell interactions (CCIs) and cell-cell communication (CCC) are critical for maintaining complex biological systems. The availability of single cell RNA sequencing (scRNA-seq) data opens new avenues for deciphering CCIs and CCCs through identifying ligand-receptor (LR) gene interactions between cells. However, most methods were developed to examine the LR interactions of individual pairs of genes. Here, we propose a novel approach named LR hunting which first uses random forests-based data imputation technique to link the data between different cell types. To guarantee the robustness of the data imputation procedure, we repeat the computation procedures multiple times to generate aggregated imputed minimal depth index (IMDI). Next, we identify significant LR interactions among all combinations of LR pairs simultaneously using unsupervised random forests. We demonstrated LR hunting can recover biological meaningful CCIs using a mouse CITE-seq dataset and a triple-negative breast cancer scRNA-seq dataset.

1 INTRODUCTION

In recent years, single-cell RNA sequencing (scRNA-seq) has been widely applied to measure gene expression at single-cell resolution, and has become a powerful tool to detect common and rare cell subpopulations, construct cell lineage and pseudotime, and identify spatial gene expression pattern, etc. While there still are many open problems and challenges remaining, scRNA-seq data analysis can be further expanded and developed to fully utilized the data for better understanding the

cell heterogeneity and gene expression stochasticity (Lahnemann et al., 2020).

Cell-cell interactions (CCIs) and cell-cell communication (CCC) are crucial for cell development, tissue homeostasis, and immune interactions in multicellular organisms (Armingol et al., 2021). In the case of cancer, tumor cells can reprogram their microenvironment to turn neutral or anti-tumor cells into tumor supportive elements (Hanahan and Weinberg, 2011;Junttila and de Sauvage, 2013), partly through secreted ligand and cell surface receptor physical interactions (Ramilowski et al., 2015). The availability of scRNA-seq data provides the great opportunities to decipher the CCIs and CCC through ligand-receptor (LR) gene expressions (Shao et al., 2020;Liu et al., 2021). Several analysis tools have been developed to infer CCC by modeling the LR co-expression data including Spearman correlation between LRs (Zhou et al., 2017;Cohen et al., 2018), product-based score from gene expression of LR pair (Kumar et al., 2018;Cabello-Aguilar et al., 2020;Hu et al., 2021), differential gene combinations (Tyler et al., 2019;Cillo et al., 2020), gene expression permutation test (Efremova et al., 2020;Dries et al., 2021;Noel et al., 2021).

Most available CCC analysis methods quantify each LR pair separately. However, biologically CCIs and CCC happen in much more complicated scenarios. In particular, the multiple ligands can compete with each other for binding on the same receptor. Therefore, the LR relationships may not be one-to-one, but would be many-to-one or many-to-many instead. To better capture the complex relationships between LR interactions, here we propose a new multivariate CCC analysis approach based on random forests (RF), which incorporates the correlations and interactions among intercellular networks to rank and prioritize the LR interactions.

2 METHOD

2.1 LR Hunting Modeling

We present a machine learning framework for LR interaction discovery, which can be used to analyze any curated LR database such as FANTOM5 (Ramilowski et al., 2015), IUPHAR (Harding et al., 2018), DLRP (Graeber and Eisenberg, 2001), or CellPhoneDB (Efremova et al., 2020).

2.1.1 Gene Expression Data Imputation

To identify LR interactions between two cell types using LR hunting analysis, we need to build the complete pseudo gene expression data matrix since ligand genes and receptor genes are from different cell types in the "interaction space" (Figure 1A). We assume that the gene expressions between two cell types follow a multivariate distribution $\mathcal{P}$ so that all the gene expression can be observed or imputed in the same framework. Formally, denote $X^{(A)}$ as an $n_A \times p_A$ matrix that records ligands gene expression for cell type A and let $X^{(B)}$ be an $n_B \times p_B$ matrix that records receptor gene expressions for cell type B. Our goal is to obtain an $(n_A + n_B) \times (p_A + p_B)$ matrix $X \sim \mathcal{P}$ so that gene associations or interactions between cell types A and B can be computed using multivariate approaches. If we are interested in the interactions between ligand genes from cell type B and receptor genes from cell type A, imputation procedure can be performed similarly as we illustrated in Figure 1A.

To this end, we applied a machine learning model, the RF missing data imputation algorithm developed by Tang and Ishwaran (Tang and Ishwaran, 2017), which was shown to be as an efficient multivariate imputation approach for high-dimensional genomic data. The RF technology is related to recursive partitioning and regression tree analyses. A single tree is inherently unstable, hence a forest of trees is "grown" from bootstrap samples of the original dataset, where an average of 37% of the data will not be sampled, referred as out-of-bag (OOB) data. The forest permits an ensemble average to be calculated across the individual trees (Breiman, 2001). We adopted the unsupervised splitting rule, where a random set of q variables, say $X_1, \ldots, X_q$, is selected to be the multivariate

pseudo-predictors. Let s be a proposed split for a pseudo-predictor X_i that splits the node t into left and right daughter nodes $t_L = \{X_i \leq s\}$ and $t_R = \{X_i > s\}$. For continuous variables, the best split is to minimize the split-statistic

$$D_q(s,t) = \sum_{k=1}^{q} \left\{ \sum_{j \in t_L} \left(X_{j,k} - \bar{X}_{t_{L_k}} \right)^2 + \sum_{j \in t_R} \left(X_{j,k} - \bar{X}_{t_{R_k}} \right)^2 \right\},$$

where $\bar{X}_{t_{L_k}}$ and $\bar{X}_{t_{R_k}}$ are the sample means of the kth pseudo response coordinate in the left and right daughter nodes. The imputation utilized the above multivariate unsupervised splitting rule for each tree where missing values are first discarded. After the forest is grown, missing data are imputed using OOB non-missing terminal node data.

2.1.2 Unsupervised Random Forests Minimal Depth Index

In order to detect LR interactions in a multivariate fashion, we adopted the unsupervised RF approach to analyze the imputed data (Shi and Horvath, 2006;Mantero and Ishwaran, 2021). RF is a modern machine learning technique that permits exploration of complex, nonlinear interrelationships (Breiman, 2001;Chen and Ishwaran, 2012). Its extension to an unsupervised algorithm composes two steps. The first step involves generating a synthetic dataset by drawing an equal number of observations from the corresponding predictor variable marginal distributions. The second step utilizes a multivariate random forest to predictthe synthetic features so that multivariate impurity splitting is able to applied in a supervised fashion.

Although the unsupervised RF can be used to cluster cells, we are more interested in selecting genes that interact with each other. We applied the minimal depth index to evaluate LR interactions in RF models (Ishwaran et al., 2010;Ishwaran et al., 2011;Chen and Ishwaran, 2013). With forests, one often observes informative variables tending to split close to the root node, where the closeness is measured by minimal depth. When considering a maximal v-subtree (Ishwaran, 2007), we could use the minimal depth of variable w to quantify the interaction between variables v and w. To illustrate this, we denote T as a random tree and define T_v a v–subtree in T for any variable v if the root node of T_v is split using v. We call T_v a maximal v–subtree if T_v is not a subtree of a larger v–subtree and define the minimal depth statistics of v, denoted by D_v, as the distance from the root node of T to the root of the closest maximal v-subtree. For example, there are two maximal v-subtrees in Figure 1B, marked in red. The maximal v-subtree on the left side is with terminal nodes 1 and 2; that on the right side is with terminal nodes 3, 4, 5, and 6.

We denote the maximal w–subtree in T_v as $T_{v,w}$ is w is used for the daughter nodes of T_v and $T_{v,w}$ is not a subtree of a larger w–subtree in T_v. The minimal depth from v to w in T_v equals to the distance from the root node of T_v to the root of the closest maximal w–subtree $T_{v,w}$, which is denoted as $D_{v,w}$. Let m be the depth of subtree $T_{v,w}$ and let l be the depth of the entire tree T. Assuming v and w are weak variables and independent with each other, we have

$$\mathbb{P}\left(D_{v,w} = d\right) = \sum_{m=d}^{l} \mathbb{P}(D_v = l - m)\mathbb{P}(D_w = l - m + d). \quad (1)$$

It was deducted that $\mathbb{P}(D_v = s) = (1 - 1/p)^{2^s - 1}\left[1 - (1 - 1/p)^{2^s}\right]$, which makes equation (1) a complicated function of d and l (Ishwaran, 2007). From this, we can normalize $D_{v,w}$ using the cumulative distribution function $\mathbb{P}\left(D_{v,w} \leq d\right)$ to evaluate LR interactions. A simpler way to normalize $D_{v,w}$ is d/m, which gives similar ranks for interactions according to empirical results.

As illustrated by Figure 1B, the interaction between variables v and w is marked with pink background: when these two variables interact with each other, we expect this depth to be smaller and this close split pattern to be repeated frequently among different trees. A single tree can be used to calculate multiple minimal depths of variables in multiple maximal subtrees, such as variables h and v in Figure 1B, where the maximal h-subtree is the entire tree. The minimal depth $D_{v,w} = d$ is

normalized by the depth of the corresponding subtree as d/m and normalized values from different maximal v-subtrees are averaged across the entire forest. We could detect variable interactions in a multivariate way adopting this imputed minimal depth index (IMDI), which averages the normalized $D_{v,w}$ and $D_{v,w}$. This normalized index ranges from 0 to 1 and smaller values indicate stronger interaction effects.

To enable the imputed dataset robustly represents the underlining distribution $\mathcal{P}$, we adopt the idea of multiple imputation, a general approach to allow for the uncertainty about the missing data by creating several different plausible imputed datasets and combining results obtained from each imputed dataset (Harel and Zhou, 2007;Carpenter and Kenward, 2014). Specifically, we generate imputed dataset $\mathbf{X}_m, m = 1, \ldots, M$, from our RF data imputation procedure described in the previous section, and use the generated IMDI, denoted by $I_{(m)}(S)$to identify interaction for gene pair S across imputed dataset. We define the aggregated IMDI for gene pair S as

$$I(S) = \frac{1}{M}\sum_{m=1}^{M} I_{(m)}(S).$$

There are $p_A \times p_B$ pair of potential interactions calculated, and we use the empirical distribution of $I(S)$ from these pairs to determine the threshold of significant interactions. The whole procedure to calculate $I(S)$ is illustrated in Figure 1C. We tested replication number m from 5, 10, 20, 50, 100, to 200 and found that the aggregated IMDI index was stable after 20 replications. We used 20 imputed datasets and aggregated those 20 IMDI for the analysis in the Section 3.

RF hunting was implemented in the open-source R software using the `randomForestSRC`. From the `randomForestSRC` R package, the function `rfsrc` was used for data imputation under default setting with 1000 trees except we set `na.action="na.impute"`; then minimal depth indices were estimated using the function `find.interaction` with method `maxsubtree`. LR hunting analysis code is available is at https://github.com/TransBioInfoLab/LRinteractions.

2.2 Pre-processing and Normalization of scRNA-seq Dataset

Two scRNA-seq datasets were used to illustrate the LR hunting approach. The first dataset is a high-quality cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) of murine spleen containing 7,097 cells with more than 1200 mRNA unique molecular identifiers (UMIs) (Govek et al., 2021). Another dataset is scRNA-seq data from five primary triple-negative breast cancer (TNBC) including 24,271 cells and 6,125 UMI detected per cell (Wu et al., 2020). For both datasets, the `SCTransform` function from the R package `Seurat_3.1.0` was used for scRNA-seq data normalization before applying LR hunting algorithm. The `CellAssign` was applied to annotate cell type for murine spleen CITE-seq data (Zhang et al., 2019).

2.3 scRNA visualization

A Seruat object was created *(*`CreateSeuratObject`, min.cells = 3, min.features = 200) witthe R package Seruat (version 3.2.3) (Stuart et al., 2019) from logNormalized scRNA from five TNBC tumors.. For clustering, the following parameters were used: `RunPCA`; `RunUMAP`, dims = 1:30; `FindNeighbors`, (dims = 1:30); and `FindClusters`. UMAP plots were generated and colored by expression levels of cell lineage markers to identify cell populations and interactions. Individual cells were plotted using previously published cell types or expression of interesting ligand-receptor pairs (Wu et al., 2020).

2.4 Circos plot visualization of ligand-receptor interaction

To summarize interactions among cell types, individual gene pair ranks were summed across the five individual patients. LR interactions were visualized with circos plots colored by interaction strength (rank sum) and line thickness representing the frequency of interaction across the tumors. Arrows indicate direction of ligand to receptor pair between cell types. Circos plots were generated using the R package `circlize` (Gu et al., 2014).

3 RESULTS

3.1 LR Hunting Recovered the Validated Cell-Cell Interactions Using scRNA-seq Data

The new digital image technologies and pipelines for multiplexed immunohistochemistry (mIHC) such as CO-Detection by indexing (CODEX) can quantify the antigens at the single-cell level to characterize tissue spatial architecture (Goltsev et al., 2018). A very recent new analysis method, spatially-resolved transcriptomics via epitope anchoring (STvEA), can integrate the CITE-seq data with mIHC images to achieve high-resolution of annotation for cell populations in the mIHC data to uncover the spatial transcription patterns (Govek et al., 2021). STvEA integrated CITE-seq and CODEX information to identify the LR pairs, thus the results are reliable and accurate. We applied LR hunting approach to only the scRNA-seq data from murine spleen CITE-seq data and then compared our results with those obtained using STvEA.

More specifically, we focused on three spatially colocalized cell populations including monocyte-derived macrophages, red-pulp macrophages, and neutrophils. We followed the procedures and LR annotations described in Govek et al. (2021). First, the mouse gene symbols were converted to the human ortholog symbols using the Bioconductor package *biomaRT*. The CellPhoneDB database was used for LR annotations (Efremova et al., 2020). Multi-subunit LR complexes were not used in this analysis due to the difficulty of annotation. The identified LR pairs by LR hunting were then converted back to their mouse orthologs to create the ranking lists.

The comparison of LR hunting results with those based on STvEA showed that LR hunting was able to detect many STvEA validated LR interactions such as monocyte-derived macrophages and neutrophils (Anxa1-Fpr1, Anxa1-Fpr2), red-pulp macrophages and neutrophils (Hebp1, Fpr2), and others (Supplementary Table 1). STvEA integrated CITE-seq and CODEX information to identify the LR pairs, thus the results are reliable. LR hunting method was able to find those validated LR pairs without borrowing the spatially expressed protein information.

3.2 LR Hunting Identified Immune, epithelial and stroma interactions in TNBC

TNBC is a diverse disease with both tumor (Lehmann et al., 2011) and stromal heterogeneity (Wu et al., 2020). Stromal-immune interactions can alter immune cell function (Gruosso et al., 2019). We applied the LR hunting approach to scRNA-seq data from five TNBC tumors to identify LR interactions between myeloid cells and either CD4 T helper (Th) cells or regulatory T cells (Treg) (Figure 2A). Professional antigen-presenting cells (APCs) such as macrophages, B cells and dendritic cells, present foreign antigens loaded on MHC-II to CD4+ Th cells. To fully activate, Th cells require a second interaction between the co-stimulatory CD80/CD86 ligands expressed on APCs and the CD28 receptor on CD4+ T cells (Figure 2B). In addition, CD4+ cells can also be converted to regulatory T cells (Treg) through consumption of IL-2 or other inhibitory cytokines, such as transforming growth factor beta (TGF-β), IL-10, and IL-35. Once converted, Tregs can interact with APCs through the immune checkpoint CTLA-4 interacting with CD80/86, impairing APCs function (Figure 2B). Using

our approach, we identified several known interactions between CD4 Th cells and myeloid APC cells, such as the costimulatory CD28-CD86 interaction, the immune activating myeloid secreted interferon gamma (IFNG) with IFNGR1/2 on CD4 cells and CD40LG-CD40 (Figure 2B and 2C). Furthermore, we were able to identify inhibitory interactions between myeloid and Treg such as CTLA4 on Tregs interacting with either CD80 or CD86, BTLA on T-regs interacting with TNFRS14 on APCs, secreted IL10 binding to the IL10RA on T cells and secreted CSF1 interacting with CSF1R on APC cells (Figure 2B and 2D). Examination of scRNA expression show that CD4-myeloid cell interactions (CD28-CD86 and CD40LG-CD40) and Treg-myeloid interactions (CTLA4-CD86 and CSF1-CSF1R) are expressed in appropriate cell types (Figure 2E and 2F).

Mammary glands consist of two differentiated epithelial cell types organized into an inner layer of luminal epithelial and an outer layer of myoepithelial cells in direct contact with the basement membrane. To better understand the directional signaling events between these two cell types in TNBC, we applied the LR hunting approach to identify interactions between luminal and myoepithelial cells (Figure 3A). We identified distinct directional interactions with multiple EGFR ligands (AREG, BTC, EREG) with the epidermal growth factor receptor (EGFR) on myoepithelial cells (Figure 3B and 3C). This signaling is consistent with the previously observed higher expression of EGFR in myoepithelial cells and that overexpression of EGFR can drive cells toward a myoepithelial phenotype in 3D culture (Ingthorsson et al., 2016). We also overserved several ligands (HBEGF and NRG1) interacting with multiple human epidermal growth factor receptors (ERBB2, ERBB3 and ERBB4) expressed on luminal cells (Figure 3B and 3C). In addition, we identified JAG1 ligand on myoepithelial cells interacting with either NOTCH2 or NOTCH3 on luminal epithelial cells, consistent with others observing NOTCH3 expression in luminal epithelial cells and JAG1 expression in the surrounding myoepithelial layer (Reedijk et al., 2005). Together these interactions describe complex multiple ligand-receptor interactions that occur between two mammary epithelial cell types.

Cancer-associated fibroblasts (CAFs) are a major component of the tumor microenvironment and can augment many characteristics of carcinogenesis including extracellular matrix remodeling, angiogenesis, cancer cell proliferation, invasion and inflammation. Two distinct populations of CAFs have been recently described in scRNA: one with features of myofibroblasts (myCAFs) and the other characterized by high expression of growth factors and immunomodulatory molecules (iCAFs) (Wu et al., 2020). To better understand how myeloid cells interact with CAFs, we applied our LR hunting approach between myeloid and either iCAF or myCAF cells (Figure 4A). We compared the interactions identified between each and show that 60% of the interactions are shared between iCAF and myCAF cells with myeloid cells (Figure 4B). Gene ontology pathway analysis interactions present in myCAFs enriched for extracellular matrix, integrin and focal adhesion (Figure 4C). However, the top pathways enriched in iCAF interactions were immune related (cytokine signaling and signaling by interleukins) in addition to extracellular matrix, focal adhesion and integrin pathways. Further examination of signaling between either iCAF or myCAF to myeloid cells revealed that myCAFs were interacting more as ligands to myeloid cells (Figure 4D and E). However, the opposite was true for myeloid ligands, in which the majority of the interactions occurred between iCAFs (Figure 4 F and 4G). Therefore, myCAFs appear to signal to myeloid cells, whereas myeloid cells provide ligands to iCAFs and the presence or absence of myeloid cells may lead to differential activation of iCAFs (Figure 4H).

For all TNBC cell pairs analyzed above, we also compared our LR hunting method with another well-known method SingleCellSignalR (Cabello-Aguilar et al., 2020). SingleCellSignalR utilizes LRscore, which is a penalized LR expression product, to rank the LR pairs. We compared results of CD4-myeloid interactions between our methods with the SingleCellSignalR method. The rankings of results by these two methods were strongly correlated (0.90-0.96) (Figure S1). The top 25 interactions

agreed well (~60% were identified by both methods), however 20% of the interactions were identified by only one method. Most of the unique interactions identified by SingleCellSignalR involved B2M and TCR interactions, while the LR hunting method identified additional key interactions (CCL5-CCR1, LGALS1-PTPRC, IFNG-IFNGR2 and CD40LG-CD4), which were not identified by LR score (Figure S1). The full ranking lists of TNBC analysis using LR hunting and SingleCellSignalR were listed in Supplementary Table 2 and 3, respectively.

4 DISCUSSION

We analyzed scRNA-seq data in a multivariate framework to identify the complex interactions between genes in different cell types and the gene pairs that are most significantly associated with each other. Traditional approaches conduct modelling of each individual LR pair without considering the correlation and high-order interaction patterns in single-cell gene expression data. To analyze the high dimensional scRNA-seq data, we first leveraged information from known LR gene pairs to filter the genes, and then used nonparametric RF approaches which had flexible statistical assumptions for the distribution of gene expression levels and nonlinear dependence of gene pairs. The merit of this approach is that after accounting for correlations and interactions multivariately, the discoveries of interacted gene pairs could be more consistent and reproducible. To account for unequal cell type distributions in different samples, we also implemented an approach that computed p-values for aggregated IMDI scores based on empirical distributions.

Using our approach, we were able to identify known interactions between differing CD4+ T cells and myeloid cells in TNBC. We also provided evidence that the directional signaling between myCAFs and iCAFs with myeloid cells is not proportional and majority of the interactions occur in the directions from myCAFs to myeloid, and myeloid to iCAFs. One limitation of our study is that only one ligand and one receptor gene pair were analyzed together in our models. Further work is needed to model complex protein structures with multiple receptors functioning as multi-subunit complexes.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Conception and design: X.S.C
Development of methodology: M.L and X.S.C,
Data acquisition: X.S and Y.B
Analysis and interpretation: M.L, Y.S, T.C.S, A.C, X.S, Y.B, L.W, B.D.L. and X.S.C,
Writing, review, and/or revision of the manuscript: M.L, L.W, B.D.L. and X.S.C,
Study supervision: X.S.C
All authors contributed to the interpretation of the results and read and approved the manuscript.

Funding

This work was supported by the following NIH grants: R01CA200987(M.L, A.C, X.S.C), P50CA098131(B.D.L), P30CA240139 (X.S.C), RF1AG061127 (T.C.S, L.W), R21AG060459 (T.C.S, L.W)

REFERENCES

Armingol, E., Officer, A., Harismendy, O., and Lewis, N.E. (2021). Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet* 22**,** 71-88.

Breiman, L. (2001). Random forests. *Machine Learning* 45**,** 5-32.

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* 48**,** e55.

Carpenter, J.R., and Kenward, M.G. (2014). *Multiple imputation and its application.*

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics* 99**,** 323-329.

Chen, X., and Ishwaran, H. (2013). Pathway hunting by random survival forests. *Bioinformatics* 29**,** 99-105.

Cillo, A.R., Kurten, C.H.L., Tabib, T., Qi, Z., Onkar, S., Wang, T., Liu, A., Duvvuri, U., Kim, S., Soose, R.J., Oesterreich, S., Chen, W., Lafyatis, R., Bruno, T.C., Ferris, R.L., and Vignali, D.a.A. (2020). Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* 52**,** 183-199 e189.

Cohen, M., Giladi, A., Gorki, A.D., Solodkin, D.G., Zada, M., Hladik, A., Miklosi, A., Salame, T.M., Halpern, K.B., David, E., Itzkovitz, S., Harkany, T., Knapp, S., and Amit, I. (2018). Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting. *Cell* 175**,** 1031-1044 e1018.

Dries, R., Zhu, Q., Dong, R., Eng, C.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., George, R.E., Pierson, N., Cai, L., and Yuan, G.C. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 22**,** 78.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* 15**,** 1484-1506.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 174**,** 968-981 e915.

Govek, K.W., Troisi, E.C., Miao, Z., Aubin, R.G., Woodhouse, S., and Camara, P.G. (2021). Single-cell transcriptomic analysis of mIHC images via antigen mapping. *Sci Adv* 7.

Graeber, T.G., and Eisenberg, D. (2001). Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat Genet* 29**,** 295-300.

Gruosso, T., Gigoux, M., Manem, V.S.K., Bertos, N., Zuo, D., Perlitch, I., Saleh, S.M.I., Zhao, H., Souleimanova, M., Johnson, R.M., Monette, A., Ramos, V.M., Hallett, M.T., Stagg, J., Lapointe, R., Omeroglu, A., Meterissian, S., Buisseret, L., Van Den Eynden, G., Salgado, R.,

Guiot, M.C., Haibe-Kains, B., and Park, M. (2019). Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. *J Clin Invest* 129**,** 1785-1800.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30**,** 2811-2812.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell* 144**,** 646-674.

Harding, S.D., Sharman, J.L., Faccenda, E., Southan, C., Pawson, A.J., Ireland, S., Gray, A.J.G., Bruce, L., Alexander, S.P.H., Anderton, S., Bryant, C., Davenport, A.P., Doerig, C., Fabbro, D., Levi-Schaffer, F., Spedding, M., Davies, J.A., and Nc, I. (2018). The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res* 46**,** D1091-D1106.

Harel, O., and Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Stat Med* 26**,** 3057-3077.

Hu, Y., Peng, T., Gao, L., and Tan, K. (2021). CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv* 7.

Ingthorsson, S., Andersen, K., Hilmarsdottir, B., Maelandsmo, G.M., Magnusson, M.K., and Gudjonsson, T. (2016). HER2 induced EMT and tumorigenicity in breast epithelial progenitor cells is inhibited by coexpression of EGFR. *Oncogene* 35**,** 4244-4255.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* 1**,** 519-537.

Ishwaran, H., Kogalur, U.B., Chen, X., and Minn, A.J. (2011). Random survival forests for high-dimensional data. *Statistical analysis and data mining* 4**,** 115-132.

Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., and Lauer, M.S. (2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* 105**,** 205-217.

Junttila, M.R., and De Sauvage, F.J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* 501**,** 346-354.

Kumar, M.P., Du, J., Lagoudas, G., Jiao, Y., Sawyer, A., Drummond, D.C., Lauffenburger, D.A., and Raue, A. (2018). Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep* 25**,** 1458-1468 e1454.

Lahnemann, D., Koster, J., Szczurek, E., Mccarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.H., Lelieveldt, B.P.F., Mandoiu, Ii, Marioni, J.C., Marschall, T., Molder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J., Saliba, A.E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., Mchardy, A.C., Raphael, B.J., Shah, S.P., and Schonhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biol* 21**,** 31.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* 121**,** 2750-2767.

Liu, J., Fan, Z., Zhao, W., and Zhou, X. (2021). Machine Intelligence in Single-Cell Data Analysis: Advances and New Challenges. *Front Genet* 12**,** 655536.

Mantero, A., and Ishwaran, H. (2021). Unsupervised random forests. *Stat Anal Data Min* 14**,** 144-167.

Noel, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F., and Soumelis, V. (2021). Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat Commun* 12**,** 1089.

Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., and Forrest, A.R. (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* 6**,** 7866.

Reedijk, M., Odorcic, S., Chang, L., Zhang, H., Miller, N., Mccready, D.R., Lockwood, G., and Egan, S.E. (2005). High-level coexpression of JAG1 and NOTCH1 is observed in human breast cancer and is associated with poor overall survival. *Cancer Res* 65**,** 8530-8537.

Shao, X., Lu, X., Liao, J., Chen, H., and Fan, X. (2020). New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell* 11**,** 866-880.

Shi, T., and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* 15**,** 118-138.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177**,** 1888-1902 e1821.

Tang, F., and Ishwaran, H. (2017). Random Forest Missing Data Algorithms. *Stat Anal Data Min* 10**,** 363-377.

Tyler, S.R., Rotti, P.G., Sun, X., Yi, Y., Xie, W., Winter, M.C., Flamme-Wiese, M.J., Tucker, B.A., Mullins, R.F., Norris, A.W., and Engelhardt, J.F. (2019). PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq. *Cell Rep* 26**,** 1951-1964 e1958.

Wu, S.Z., Roden, D.L., Wang, C., Holliday, H., Harvey, K., Cazet, A.S., Murphy, K.J., Pereira, B., Al-Eryani, G., Bartonicek, N., Hou, R., Torpy, J.R., Junankar, S., Chan, C.L., Lam, C.E., Hui, M.N., Gluch, L., Beith, J., Parker, A., Robbins, E., Segara, D., Mak, C., Cooper, C., Warrier, S., Forrest, A., Powell, J., O'toole, S., Cox, T.R., Timpson, P., Lim, E., Liu, X.S., and Swarbrick, A. (2020). Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *EMBO J* 39**,** e104063.

Zhang, A.W., O'flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., Mcpherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B., Lai, D., Mottok, A., Sarkozy, C., Chong, L., Aoki, T., Wang, X., Weng, A.P., Mcalpine, J.N., Aparicio, S., Steidl, C., Campbell, K.R., and Shah, S.P. (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 16**,** 1007-1015.

Zhou, J.X., Taramelli, R., Pedrini, E., Knijnenburg, T., and Huang, S. (2017). Extracting Intercellular Signaling Network of Cancer Tissues using Ligand-Receptor Expression Patterns from Whole-tumor and Single-cell Transcriptomes. *Sci Rep* 7**,** 8815.

Figu

Figure Legends

Figure 1. Illustration of RF Methods. (A) Data sheet and imputation illustration. (B) The minimal depth of *w* in a maximal *v*-subtree. Letters in parent nodes identify the variable used to split the node. There are two maximal *v*-subtrees, marked in red. The maximal *v*-subtree on the left side is with terminal nodes 1 and 2; that on the right side is with terminal nodes 3, 4, 5, and 6. The Minimal depth of *w* in the second maximal *v*-subtree is the depth of *w* (*d*=2 marked with pink background) normalized by the subtree depth (*m*=3), which is *d/m* =2/3. (C) Model workflow for LR hunting.

Figure 2. Differential LR interactions between myeloid cells and either Th or Treg cells. (A) UMAP plot shows distinct cell populations in five TNBC tumors. (B) Image shows known interactions identified by LR Hunting between myeloid antigen presenting cells and either CD4 Th or T-regs. Colored arrows indicate direction of signaling from ligand to receptor for stimulatory (green) and inhibitory (red) events. Dashed arrows indicated secreted ligands. Circos plots show the top interactions and by direction for (C) myeloid and CD4 T cells and (D) myeloid and T-reg cells. LR interactions are colored by interaction strength (rank sum) and line thickness represents the frequency of interaction across the five tumors. (E) UMAP plots show expression of CD4 Th and myeloid cell markers and expression of LR interactions for CD28-CD86 and CD40L-CD40. (F) UMAP plots show expression of T-reg and myeloid cell markers and expression of LR interactions for CTLA4-CD86 and CSF1-CSF1R.

Figure 3. Multiple ligand receptor interactions between luminal epithelial cells and myoepithelial cells. (A) UMAP plot shows distinct cell populations in five TNBC tumors. (B) Image shows unique ligand (red) receptor (black) interactions between myoepithelial and luminal breast cells. (C) Circos plots show the top interactions and by direction between myoepithelial and luminal cells. LR interactions are colored by interaction strength (rank sum) and line thickness represents the frequency of interaction across the five tumors.

Figure 4. Direction of LR interactions favor myofibroblast CAFs to myeloid and myeloid to inflammatory CAFs. (A) UMAP plot shows distinct cell populations in five TNBC tumors with cells of interest in bold. (B) Venn diagram of interactions between myeloid and myCAF with interaction between myeloid and iCAF. (C) Gene ontology pathway analysis (C2 canonical pathways) of differential interactions unique to myCAFs (pink) or iCAFs (blue). (D) Circos plots show the top interactions between iCAF and myeloid cells (iCAF>myeloid) or myCAF to myeloid (myCAF>myeloid). (E) Venn diagram shows overlap of interactions between CAFs and myeloid cells. (F) Circos plots show the top interactions between myeloid cells and iCAFs (iCAF>myeloid) or myCAFs(myCAF>myeloid). (E) Venn diagram shows overlap of interactions between myeloid to CAF interactions. (H) Summary of directional interactions between myeloid cells and myCAFs or iCAFs.

Figure 1.JPEG

Figure 1

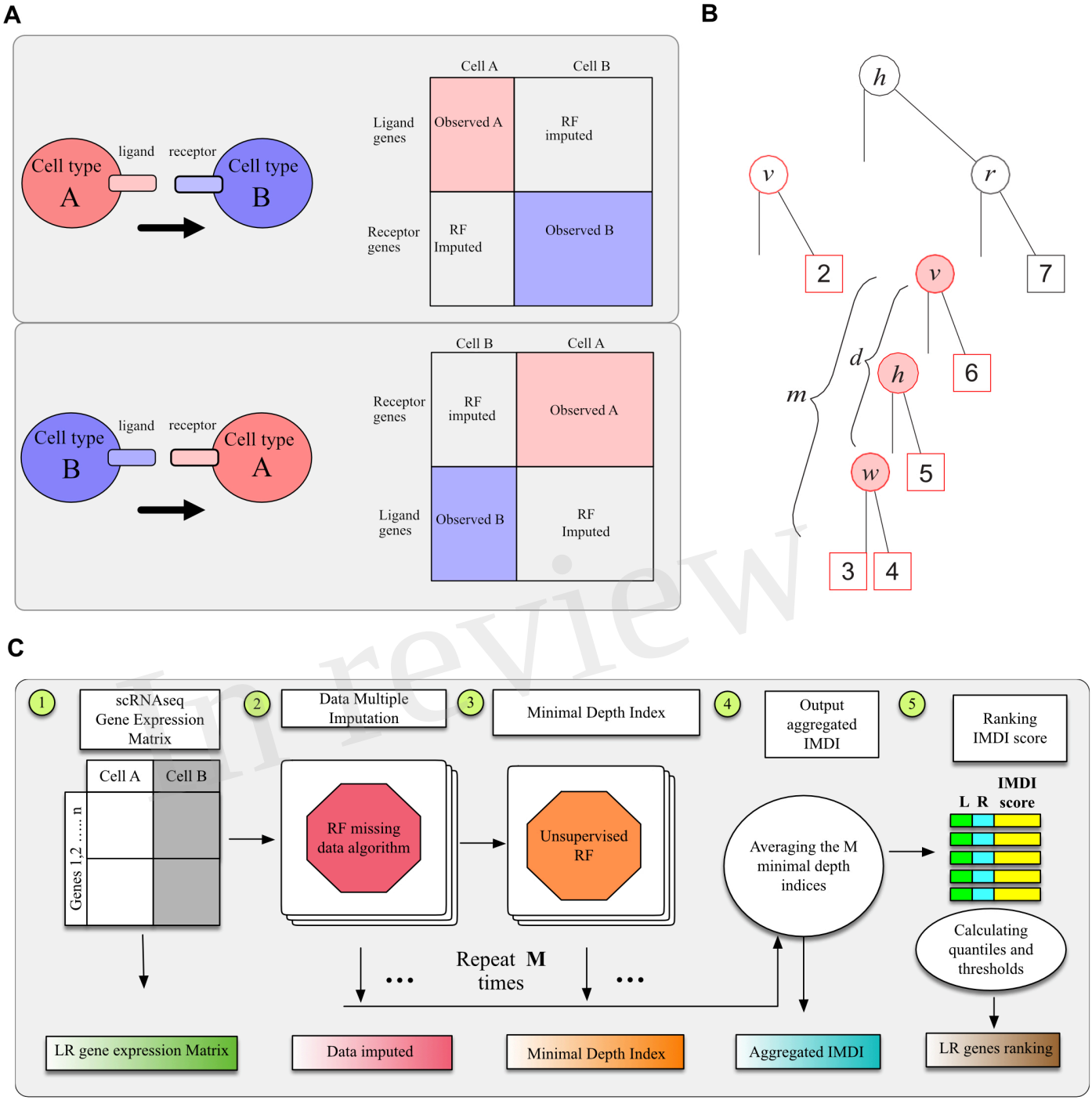

Figure 2.JPEG

Figure 2

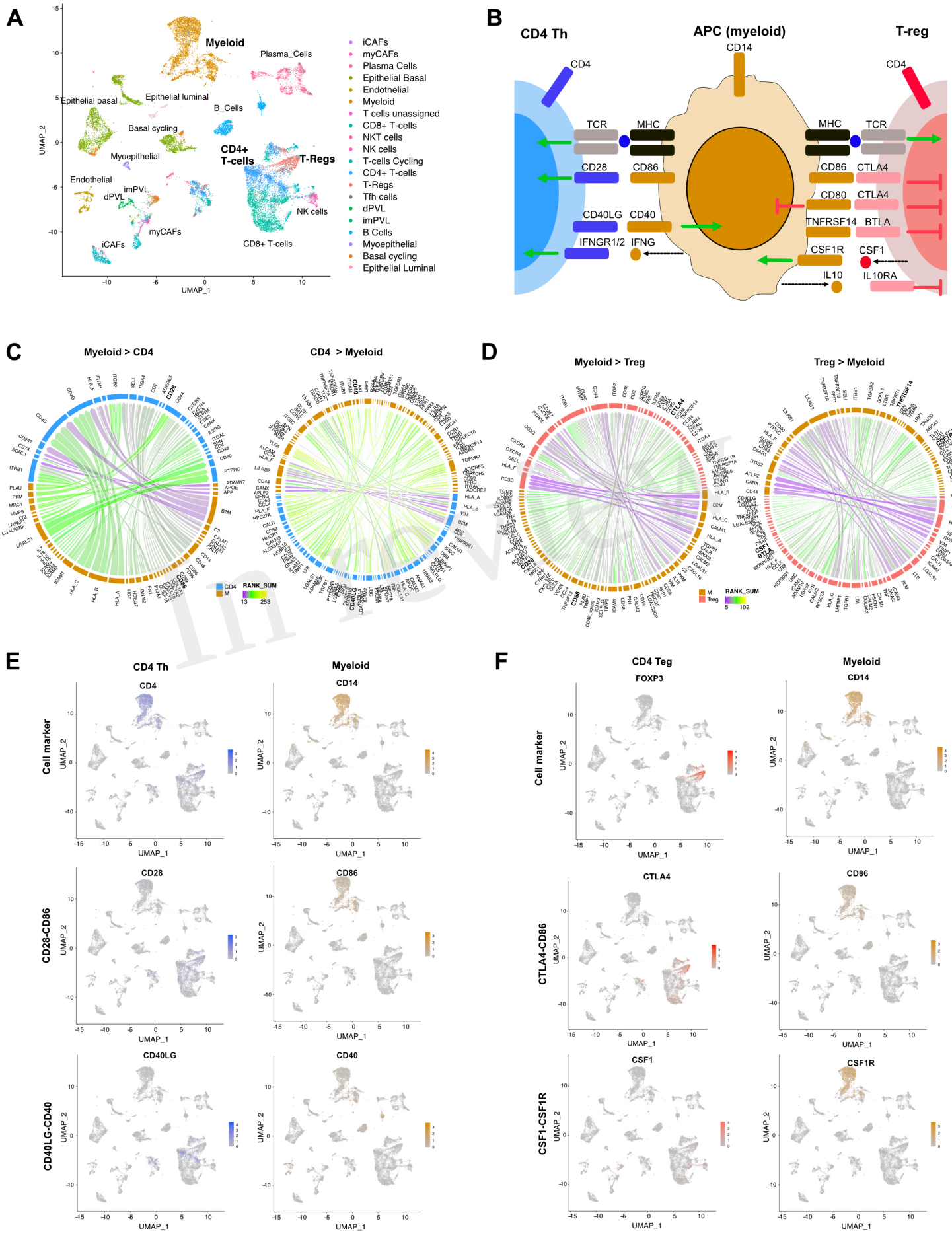

Figure 3.JPEG

Figure 3

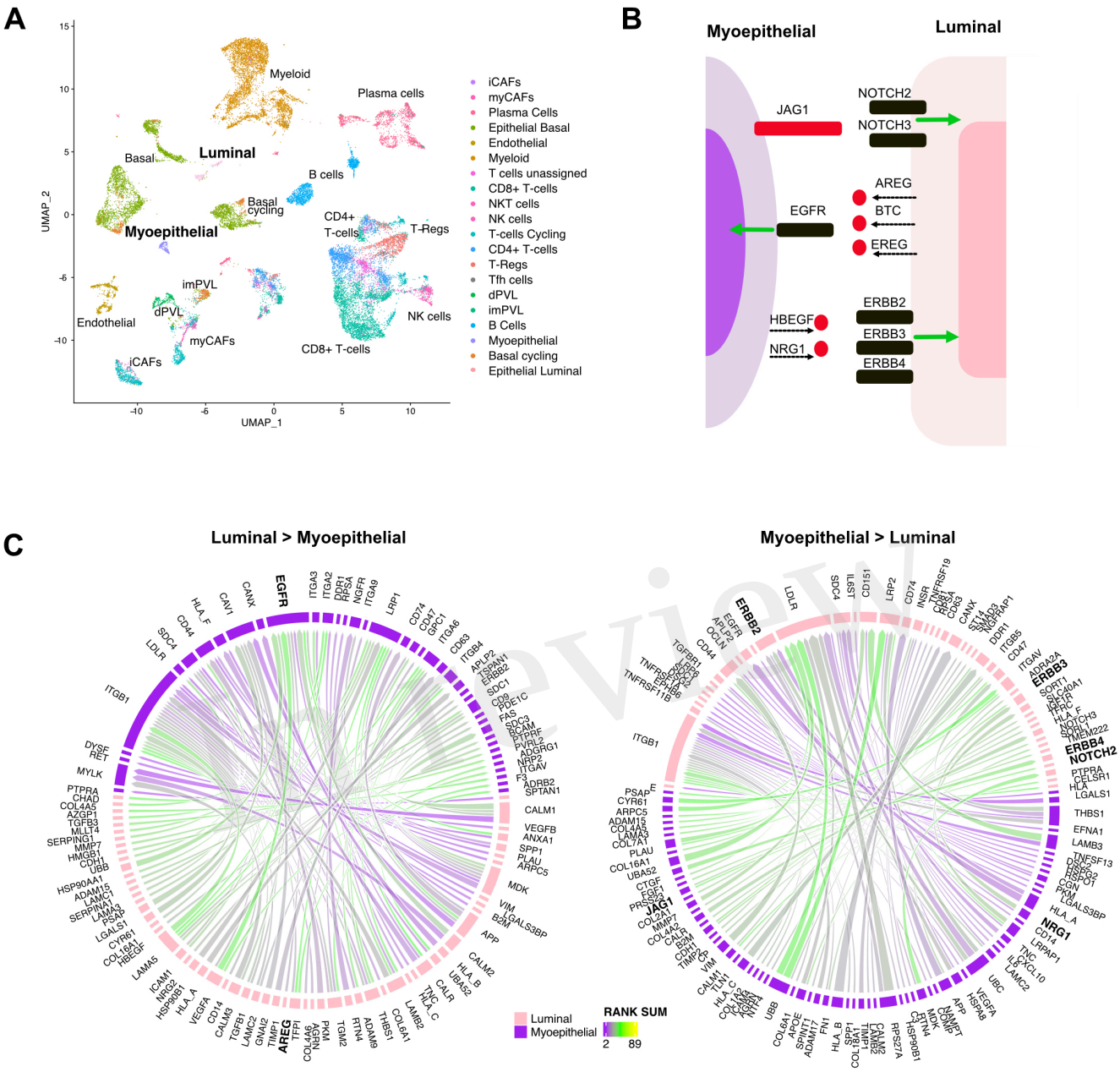

Figure 4.JPEG

Figure 4