Annual Report for Cell Discovery Network -- Rongbin Zheng

The proposal aimed to investigate cell-cell communication (CCC) by developing computational methods for phenotype-associated CCC network analysis and analyzing scRNA-seq data from the embryonic heart with left ventricular noncompaction (LVNC). CCC is crucial for maintaining the functions and homeostasis of cells, organs, and intact systems. It can be mediated by various types of molecules, e.g., proteins and metabolites. Single-cell RNA sequencing (scRNA-seq) data provides opportunities to model CCC networks in tissues. However, metabolite-mediated CCC (mCCC) analysis using scRNA-seq has been challenging.

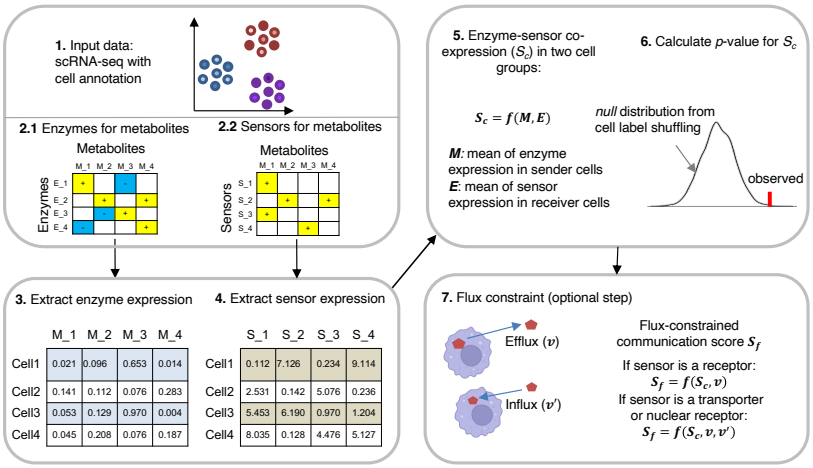

Figure 1. Cartoons to show the seven steps in MEBOCOST to identify cell-cell metabolite-sensor communications. (1) The RNA expression data and cell type annotation obtained from scRNA-seq were taken as the input data to MEBOCOST. (2) A knowledgebase of enzymes (2.1) and sensor proteins (2.2) for metabolites is incorporated into the MEBOCOST algorithm. (3) The RNA expression values of the metabolite enzymes were extracted from scRNA-seq data. (4) The RNA expression values of sensor proteins for metabolites were extracted from the scRNA-seq data. (5) Calculate the enzyme-sensor co-expression score between two cell groups by taking the product of the mean of enzyme expression values in the sender cells and the mean of sensor expression values in the receiver cells. (6) Shuffling cell labels to generate a statistical *null* distribution to calculate the *p*-value for the enzyme-sensor co-expression score.

To address this gap, we have developed MEBOCOST to identify sender cells that secrete metabolites and receiver cells that sense these metabolites through mechanisms that include cell surface transporters, cell surface receptors, and nuclear receptors (**Figure 1A**). MEBOCOST allows prediction of mCCC for over 700 metabolite-sensor interactions. It takes scRNA-seq data as input and identifies which enzymes and sensors of a metabolite were highly expressed in the sender and receiver cells, respectively. The prediction of mCCC in MEBOCOST includes 6 steps (**Figure 1B**), including (1) taking scRNA-seq data as input, (2) obtaining prior knowledge of metabolite-sensor partners and metabolite-enzyme associations, (3) metabolite presence inference, (4) extracting sensor expression, (5) calculating the communication score, and (6) calculating p-values by permutation test, (6) integrating Flux-Balance analysis (FBA) into our MEBOCOST workflow to account for metabolite secretion in sender and uptake in receiver cell types.

To enable investigation of disease- or phenotype-related mCCC, we expanded the MEBOCOST workflow to perform differential mCCC analysis. For each mCCC event, defined by a metabolite-sensor pair and sender-receiver cell types, we calculated the $\log_2$ fold change in communication

scores between disease and normal samples. To assess statistical significance, we generated background $\log_2$ fold changes using communication scores derived from randomly shuffled cell populations. These *null* distributions were then used to estimate *p*-values for the observed fold changes, followed by false discovery rate correction.

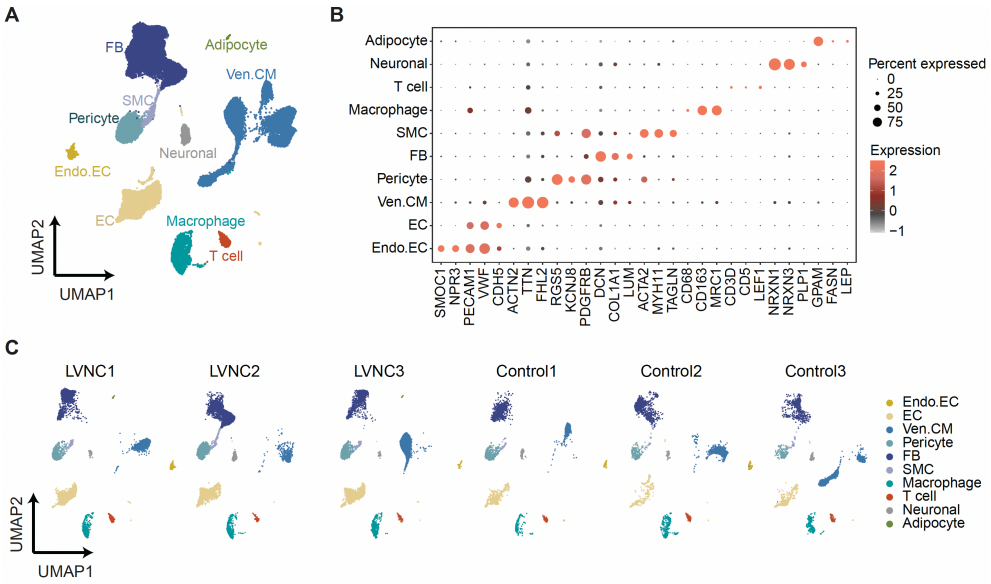

Figure 2. A. UMAP of heart scRNA-seq samples with cell type annotation. B. Dot plot showing the marker gene expression for each cell type. C. UMAP of heart scRNA-seq samples divided by different conditions. Color represents different cell type.

As another accomplishment of this proposal, we have successfully generated scRNA-seq samples from the mouse embryonic heart of control and LVNC conditions. UMAP analysis identified nine major cell populations, including fibroblasts (FB), smooth muscle cells (SMC), pericytes, endothelial cells (EC), endocardial endothelial cells (Endo.EC), macrophages, T cells, ventricular cardiomyocytes (Ven.CM), neuronal cells, and adipocytes (**Figure 2A**). Cell-type assignments were supported by the expression of established marker genes, such as COL1A1/COL3A1 in fibroblasts, ACTA2/MYH11 in SMCs, RGS5/ABCC9 in pericytes, PECAM1/VWF in endothelial cells, CD68/CD163 in macrophages, CD3D/CD3E in T cells, and RYR2/PLN in cardiomyocytes (**Figure 2B**). Separate visualization of the three LVNC and three control samples showed that all major cell populations were represented across individuals and displayed broadly comparable UMAP distributions, indicating consistent cell-type annotation and successful integration across samples (**Figure 2C**).

In the future, we will apply the MEBOCOST and its differential analysis function to investigate LVNC-related mCCC events. We will prioritize key CCC events for further experimental validation. The MEBOCOST package and its detailed tutorial can be freely accessed on GitHub: https://github.com/kaifuchenlab/MEBOCOST. This work has produced a broadly applicable bioinformatics tool that will be useful for the community to investigate cell-cell communication across many cell types in numerous other biological and disease models.