Analyzing and Improving Reference Cluster Mapping Methods for Cross-Species Data Sets and Interpreting Similar Cell Types Across Species

Chenjishi Lin, Amie Choe, Hadassah Mayerfeld (Mentors: Drs. Bianca Dumitrascu and Andrew Blumberg, Yining Liu, 2024 IICD SRP)

Single-cell RNA sequencing measures the gene expression levels for a cell. Cells with similar levels of gene expression can be grouped and categorized via clustering and the "marker approach" for cell type annotation. Reference cluster mapping can be used to label a query data set based on a reference data set by constructing a clusterwise dissimilarity matrix, creating a weighted graph to map the relationships between clusters, and calculating the minimum matching. While this method is efficient for single-species cell analysis, this method fails when considering cross-species comparison because data sets from different species are sampled on different genes and there are no one-to-one homologues.

The project will focus on two computational tools, SATURN (Species Alignment Through Unification of Rna and proteiNs) and RefCM (reference cluster mapping), to address the challenge of cross-species analysis. SATURN is a deep learning method that couples protein embeddings with RNA expression to map cells from datasets sampled on different genes to a universal "macrogene" space of functionally related genes.

RefCM uses the principles of optimal transport theory to map cell-type clusters across different tissues, sequencing methods, and species by exploring the geometric properties of data distribution in gene expression space. Optimal transport methods, like the Gromov-Wasserstein distance, allows for the alignment and alteration of scRNA-seq data in a way that minimizes the discrepancy in gene expression patterns and enables the identification and comparison of gene expression across species without requiring direct genetic correspondence.

[Chenjishi Lin]

We will analyze the robustness of optimal transport-based cell type transfer with respect to the underlying clustering methodology. RefCM presupposes that the data input by the user has undergone a preferred type of clustering algorithm on cell type labels, such as K-means. However, various clustering algorithms yield differing results when applied to optimal transport. Consequently, it is critical to ascertain the sensitivity of optimal transport and adopting to one appropriate standard clustering methodology in RefCM.

[Amie Choe]

We will then aim to compare optimal transport-based annotation transfer across species with state-of-the-art methods, including SATURN. By applying optimal transport, we seek to evaluate and understand how various methods facilitate the comparison of cell types across species.

[Hadassah Mayerfeld]

We will interpret gene clusters explaining cell type matchings. We will consider different algorithms to calculate the minimum matching and compare the gene clusters explaining the resulting cell type matchings from each algorithm. We will consider the biological significance of the gene clusters that explain the cell type matchings across all the minimum matching algorithms as well as the genes that contribute to the matching for just select algorithms.