Single-cell RNA Sequencing Data Analysis

Mojtaba Ghasemi

August 2023

1. Motivation and Problem

In recent years, the field of single-cell genomics has witnessed an unprecedented level of growth, pushing the boundaries of our biological understanding. The ability to measure and analyze individual cells' DNA, RNA, and protein constitution has allowed for an in-depth view of complex biological systems, yielding insights into processes such as human embryonic development, novel disease-associated cell types, and potential cell-targeted therapeutic interventions. Machine learning models can reveal unique characteristics of individual cells, providing insights into how gene regulation affects blood and immune cell differentiation and maturation over time.

This proposal seeks to delve further into this rich biological data by employing advanced machine-learning models to explore how these genetic factors co-vary as bone marrow stem cells mature into blood cells.

While we have come a long way in single-cell genomic research, the analysis methods for multimodal single-cell data remain insufficient, especially considering the temporal dynamics alongside state changes over time. This leaves us with a knowledge gap in comprehending how genetic information flows and regulates dynamic cellular processes such as cell differentiation and maturation. This project aims to address this by investigating how DNA, RNA, and protein measurements co-vary in single cells as bone marrow stem cells develop into more mature blood cells.

2. Potential Client

The potential clients for this project could be various stakeholders in the medical and biotechnology industry. This includes research institutions focused on genomics and cell biology, pharmaceutical companies investing in cell-targeted therapeutic interventions, and healthcare providers interested in developing novel disease treatments. Additionally, organizations like Cellarity, Chan Zuckerberg Biohub, the Chan Zuckerberg Initiative, Helmholtz Munich, and research institutions such as Yale University could also be potential clients, given their interest in single-cell analysis.

3. Data Analysis:

For this project, we will be working with a subset of a 300,000-cell time course dataset generated from CD34+ hematopoietic stem and progenitor cells (HSPC). The analysis will involve machine learning techniques that can handle the data's sparsity, noise, and variation. We will account for different feature spaces and shared and unique variations between modalities and batches.

4. Solution and Approach:

Our approach will include applying and adapting machine learning techniques to multimodal single-cell data to predict how DNA, RNA, and protein measurements co-vary in single cells as bone marrow stem cells mature into blood cells. The main objective is to reveal the unique characteristics of individual cells and provide insights into the rules governing these complex regulatory processes.

5. Deliverables:

By the end of the project, we aim to deliver a detailed report outlining our findings, along with a comprehensive machine-learning model that can predict DNA, RNA, and protein co-variance in developing cells. We also plan to present visual representations of the data to illustrate the developmental changes in the cells better. Our final deliverable will also include the Jupyter notebooks for streamlining the analysis and the model.

6. Source:

The dataset for this project comprises single-cell multiomics data collected from mobilized peripheral CD34+ hematopoietic stem and progenitor cells (HSPCs) isolated from four healthy human donors at five-time points by ALLCELLS, using two single-cell assays technology (Multiome and CITEseq) developed by 10x Genomics company.

https://allcells.com/research-grade-tissue-products/mobilized-leukopak/
https://www.10xgenomics.com/products/single-cell-multiome-atac-plus-gene-expression
https://www.kaggle.com/competitions/open-problems-multimodal/data