Comparing Models for Single-Cell RNA Sequencing Data

BIOS 6643 – FALl 2019 – Final Project

Lee Panter

Introduction

The discovery of DNA and RNA and, more importantly, the development of genomic sequencing methods has proven to be a valuable tool of scientific research. However, the methods developed in the first-generation of technology relied on "bulk" sampling, which could only estimate population-average expression of RNA and DNA. In order to obtain a more complete understanding of how the cellular landscape functions, estimates of cell-to-cell variability would need to be estimated. Development of single-cell RNA sequencing (scRNAseq) technology has increased to satisfy this need, but a need for statistical analysis is still outstanding. Previous methods used to model bulk RNA sequencing data do not account for the correlated nature of scRNAseq data.

This paper will compare six different modeling approaches on an observational scRNAseq data set obtained from a Lupus Nephritis Study of 33 patients across the United States. Two RNA genes were selected to be the predictor-response pair to simplify the modeling process. The main goal of this report is to investigate the ways in which parameter estimates vary as modeling methodology is altered. It is hoped that the results of this investigation are useful for the development of future models for scRNAseq data sets.

Data

A single-cell RNA sequencing (scRNAseq) expression profile is a matrix of count-values representing a time and space "snapshot" of the magnitude of activity of genomic features of a single cell. ("Gene Expression Profiling - Wikipedia," n.d.) In its original form, the data matrix contained 9,560 single-cell observations clustered within 27 samples (5 control not included in data). Each observation contained the expression of 38,354 genetic features. Single-Cell data is often unreliable, protocol dependent, and can often have batch effects. Data quality control (QC) filters out redundant measures, and dead cell observations. The Seurat guided tutorial (Satija and others 2018) was used to perform quality control, filtering out observations with:

- Percent Mitochondrial DNA > 60
- Genetic Features Expressed $< 1{,}000$
- Genetic Features Expressed $> 5{,}000$
- B-cells only

These quality control measures reduced the original data by 88%, leaving only 1,110 observations clustered within 15 samples. Two genes (MALAT1 and CD19) were then selected from the set of genetic features in the initial data to be studied due to a higher magnitude of correlation. MALA1 has been consistently linked with cancer metastasis, cell migration, and cell regulation. ("MALAT1 Gene - Genecards | Malat1 Rna Gene," n.d.) CD19 encodes a cell surface molecule which regulates lymphocyte proliferation and differentiation. [CD19Gene32:online].

Histograms and a joint distribution scatter plot was constructed to visualize the distributions of the selected variables (Appendix: Fig1-Fig3). The presence of zeros in the data indicated that the distribution might be well suited for a zero-inflated mixture model. Specifically, since the response was count-valued, the histograms indicated that a zero-inflated Poisson Generalized Linear Model or Generalized Linear Mixed Model would be appropriate. Additionally, while normality was not expected, log-transformations were also applied (Appendix: fig4-fig6), and resulted in approximate normality of the response MALAT1 and a bimodal distribution of the predictor CD19.

References

"Gene Expression Profiling - Wikipedia." n.d. `https://en.wikipedia.org/wiki/Gene_expression_profiling`.

"MALAT1 Gene - Genecards | Malat1 Rna Gene." n.d. `https://www.genecards.org/cgi-bin/carddisp.pl?gene=MALAT1`.

Satija, R, and others. 2018. "Seurat: Guided Clustering Tutorial." *Satija Lab Http://Satijalab. Org/Seurat/Pbmc3k_tutorial. Html*.