Comparing Models of Subject-Clustered Single-Cell Data

Version 3.0

Lee Panter

Abstract

Single-Cell RNA sequencing data represents a revolutionary shift to the bioinformatic approaches being used to decode the human transcriptome. Such data are becoming more prevalent, and are being extended to multiple individuals, enabling analysis of subject-level relationships. However, it is not clear how to conduct this subject-level analysis. Current methods do not account for nested study designs in which samples of hundreds, or thousands of cells are gathered from multiple individuals. Therefore, there is a need to outline, analyze, and compare methods for estimating subject-level relationships in single-cell expression. Here, we compare three modeling strategies for single-cell RNA sequencing expression estimation in subject-correlated study design data. Each of the three methods: Linear Regression with Fixed Effects, Linear Mixed Effects Models, and Generalized Estimating Equations will have a detailed outline presented. We then compare the regression estimates and standard errors for each modeling method using real single-cell data from a Lupus Nephritis study of 27 subjects. We hoped that this paper presents insights into modeling single-cell expression data, and aids researchers with down-stream analyses.

Introduction

Traditional methods of sequencing the human transcriptome involve anylizing expression profiles of genetic material obtained from thousands or even millions of cells. It is for this reason, that traditional methods are often reffered to as "bulk" techniques. Bulk techniques are informative regarding population-averaged parameters, but often fail to capture the variability in expression profiles within a single sample of genetic material [1].

Single-cell RNA sequencing (scRNA-seq) data sets involve anylizing the expression profile for a single cell. Analysis of such data allows for more specific information estimation, including population variability. This data has been used in research applications for identifying rare cellular subpopulations and characterizing genes that are differentially expressed across conditions [2]. Technological developments in whole-genome sequencing have made generating single-cell data more cost effective, ultimately leading to data proliferation in the form of:

- increase in the amount of data within a single population source (e.g a single person),and
- an increase in the number of distinct population sources (multiple people in one data set).

The feasability of single-cell data sets across multiple subjects motivates a need to compare, test, and integrate methods that can adequately model single-cell data and account for the correlation of repeated measures within subjects (many single-cell observations within each subject).

Code and Data

All code for the above analysis was written and evaluated in RStudio Version 1.2.1335, and is available for download at the following GitHub repository:

https://github.com/leepanter/MSproject_RBC.git

Additionally, a link to all necessarry and referrence data files (including original data) are contained in the following Google Drive:

https://drive.google.com/open?id=1gjHaMJG0Y_kPYWj5bIE4gRJU5z9R2Wqb

References

1. Macaulay IC, Voet T (2014) Single cell genomics: Advances and future perspectives. *PLoS genetics* 10: e1004126.

2. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* 17: 63.