Comparing Models of Subject-Clustered Single-Cell Data

Version 6.0-Introduction

Lee Panter

Introduction

Traditional methods of sequencing the human transcriptome involve analyzing the combined genetic material of thousands or even millions of cells. These, so called “bulk” techniques provide information about the average gene expression across the cells, but often fail to capture the underlying variability in expression profiles within the sample of cells [1].

The techniques used for single-cell analysis and the information obtained from these analyses do not suffer from the same inability to estimate expression profile variation within a sample of cells as traditional “bulk” techniques. The sampling methods employed for single-cell RNA sequencing (scRNA-seq) data acquisition obtain measurements of transcriptomic information specific to individual cells. Hundreds or even thousands of RNA-sequencing profile measurements, each specific to a single-cell, can be used to estimate estimate expression variability across the cells within the sample. This feature of single-cell data analysis is suited for research applications that seek to identify rare cellular subpopulations, or characterize expressions that are differentially expressed across conditions [2]. Additionally, technological developments have made generating single-cell data more cost effective, and easier to obtain on multiple sample-sources, most noteably on multiple individuals.

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

Here, we compare three methods for modeling scRNA-seq expression profiles that account for within-subject correlation: Linear Regression with Fixed Effects, Linear Mixed Effects Models with Random Effects, and Generalized Estimating Equations. We will present the framework for each method to reflect the fitting of a predictor-response pairing as defined by: two different Linear Regression linear predictors, two different Linear Mixed Effects linear predictors, and a single GEE linear predictor. We will assess the estimates assigned to each model for the parameter that reflects subject inspecific interaction between predictor and response (main-effect slope). This parameter will be assesed for stability across model, and across predictor-response pairings using subject-correlated single-cell data from a study of 27 Lupus Nephritis cases. We will also evaluate standard errors and test statistics for this parameter.

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.