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 cell sample 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 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 notably on multiple individuals.

The utility of single-cell data, and the feasibility 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, I compare five methods for modeling scRNA-seq expression profiles that account for within-subject correlation: linear modeling (LM), linear modeling with subjects as fixed effects (LM-FE), linear mixed effects models with subjects as random intercepts (LMM-RI) and random slopes (LMM-RS), and generalized estimating equations. I will present the framework for each method in the context of a generalized predictor-response pairing. Then I will

assess each model's estimate of the fixed effect slope parameter for stability across modeling approach 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.

Discussion

Here, I compare five modeling strategies for detecting subject level associations in single-cell RNA sequencing data gathered over 27 subjects from a Lupus Nephritis study. I compare estimates of a fixed effect slope parameter generated by five modeling techniques: linear modeling (LM), linear modeling with subjects modeled as fixed effects (LM-FE), linear mixed effects models with subjects modeled as random intercepts (LMM-RI) and random slopes (LMM-RS), and generalized estimating equations (GEE).

I find that population average models (i.e. LM and GEE) and subject specific intercept models (i.e. LMM-RI and LMM-RS) tend to produce similar results within the same model class (population average or subject specific intercept) but different results between model classes. The highest standard errors are indicated in the LMM-RS model, and the lowest standard errors in the LMM-RI model. LM-FE standard error is also found to be smaller than both LM and GEE standard error values. Nested model comparisons indicate that inclusion of subject specific terms is advisable at all levels (fixed and random, intercept and slope) with exception of the random slope in the $FBLN1 \sim CD34$ variable paring.

Interpretations of subject specific parameters are contextually authentic provided that they are used in inference conditional to their subject of origin. Conversely, interpretations of population average parameters are accurate when they are used for inference on a population's hypothetical representation of centrality. Under conditions of linearity and error normality, it can be shown that subject specific parameters are marginal representations of population average parameters. This distinction explains the parameter estimate disparities as estimated

between the LM/GEE methods compared to the LM-FE/LMM-RI methods.

This analysis is subject to several drawbacks and limitations. All the results are based on evidence obtained from just two single-cell RNA sequencing variable pairings. In the future, comparing the consistency of these models over all model pairs is needed. Additionally, single-cell RNA sequencing data is heavily influenced by protocol dependencies and measurement inconsistencies. Quality control must be carefully considered and conducted prior to any analysis.

The utility and promise of single-cell RNA sequencing data indicates that such data will become more prevalent and will be extended to multiple subject samples. I have presented an initial comparison of methods for detecting subject-level associations in single-cell RNA sequencing data sets.

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.