Comparing Models of Subject-Clustered Single-Cell Data

v1

Lee Panter

# Abstract

Single-cell RNA sequencing (scRNA-seq) represents a revolutionary shift to the analytic approaches being used to decode the human transcriptome. Single-cell data has been used to: visualize cellular subpopulations with unsupervised clustering methods, test for differential expression rates across conditions using logistic and mixture modeling, and reconstruct spatio-temporal relationships using network analysis. While these successes demonstrate the utility and promise for single-cell methods, they do not demonstrate the practical need to generalize to single-cell data over multiple individuals. This paper looks to compare three different modeling strategies for RNA-seq expression estimation for data with individual-level clusting. The modeling approaches will be compared theoretically against an Ordinary Least Squares model, and analytically motivated by data from a Lupus Nephritis study. It is hoped that this paper will present new approaches to modeling single-cell expression data, and will be useful not only for Statisticians, but also Geneticists and Microbiologists.

# Introduction

Single-cell analysis has emerged as a leading methodology for the analysis of transcriptomic data. [1] Such data sets have demonstrated their utility in research contexts for identifying rare subpopulations, characterizing genes that are differentially expressed across considtions, and in spatio-temporal inference. [2] Additionally, advances in whole genome amplification and cellular isolation techniques have made single-cell data sets more accessible and informative than ever before. [1]

Traditional methods for analylizing single-cell data for subpopulations commonly involve unsupervised clustering techniques including Principle Components Analysis (PCA) and K-nearest neighbors (KNN). These methods have been shown to be effective in identifying rare nerological cells within a homogeneous population. [3]. Such clustering methods, and additional (non-linear) methods such as the t-distributed stochastic neighborhood embedding (t-SNE) are also useful for visualizing high-dimensional data and have been used to find multi-dimensional boundary values for distinguishing heathly and cancerous bone marrow samples. [4] While both these studies involve single-cell data which incorperates multiple subjects, the modeling methodologies do not provide estimates for subject-factor effects.

# 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.

3. Ståhlberg A, Andersson D, Aurelius J, et al. (2010) Defining cell populations with single-cell gene expression profiling: Correlations and identification of astrocyte subpopulations. *Nucleic acids research* 39: e24–e24.

4. Amir E-aD, Davis KL, Tadmor MD, et al. (2013) ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology* 31: 545.