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.

Single-cell data has been used to target treatments by characterizing differential expression across condition. Model-based Analysis of Single-cell Transcriptomics (MAST) has been used to compare "primary human non-stimulated" and "cytokine-activated" mucosal-associated invariant T-cells. [5] Single-Cell Differential Expression (SCDE) was also used to compare 92 mouse embryonic fibroblasts with 92 embryonic stem cells. [6] Neither of these studies included sampling across multiple subjects, and the resulting models do not account for possible correlation within subjects that might be present.

Network modeling approaches, in conjunction with single-cell data have provided the opportunity to learn about cellular heirarchies, spatial relationships, and temporal progressions.

Weighted Gene Co-Expression Network Analysis (WGCNA) has been used to find delineations in both human and mouse embryonic transcriptome dynamics during progression from oocyte to morula. [7] A similar analysis was performed using Single-cell Clustering Using Bifurcation Analysis (SCUBA), and was verified using Reverse Transcription Polymerase Chain Reaction (RT-PCR) data over the same single-cell measurements. [8] The studies conducted using network modeling approaches targets single-cell sources at multiple time points or with distinct measures that could be compared using a pseudo-time metric. Diversification of the single-cell data by incorperating multiple subjects is not considered or adressed.

Down-stream analyses of single-cell data can be a very useful tool for transcriptome analytics. Technological advances in cellular isolation and genetic material amplification will lead to a rise in single-cell data prevalence, and a corresponding rise in the prevalence of multiple-subject single-cell data sets. Therefore, there is a clear need to develope, test and integrate alternative methods that can accurately and precisely model single-cell data and account for the correlation of repeated measures within subject samples.

This paper seeks to satisfy this need by proposing _____ methods for modeling scRNA-seq expression profiles that account for within-subject correlation differently. A motivating example consisting of scRNA-seq observations across multiple subjects with Lupus Neprhitis will be detailed. We will provide the model theory and comparisons in the context of this example. Results of the various modeling approaches will be compared and any conclusions discussed along with implications and future research following the discussion.

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.

5. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: A revolutionary tool for transcriptomics. *Nature reviews genetics* 10: 57.

6. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nature methods* 11: 740.

7. Xue Z, Huang K, Cai C, et al. (2013) Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature* 500: 593.

8. Marco E, Karp RL, Guo G, et al. (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences* 111: E5643–E5650.