Comparing Models of Subject-Clustered Single-Cell Data

Version 2.0

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 is 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 in the microbiome using network analysis. These accomplishments demonstrate the utility and promise of single-cell research; however, if numerical results are desired, each analysis needs to be altered upon the hypothetical inclusion of single-cell observations sourced from multiple individuals. Since single-cell data acquisition is increasing in efficiency and decreasing in cost, data sets featuring single-cell observations from multiple subject sources can be expected to rise in prevalence as a default method of attempting to improve analytic power. Therefore, there is a practical need to outline, analyze, and compare current methods for obtaining numerical parameter estimates for between-subject observation correlation. This paper looks to compare three different modeling strategies (each with different estimates for between-subject correlation parameters) for scRNA-seq expression estimation in data with subject-level clusting. The modeling approaches are compared theoretically, and analytically, motivated by data from a Lupus Nephritis study. It is hoped that this paper presents insights

into modeling single-cell expression data, as well as aids researchers with down-stream analyses, and future theoretical/analytic methodology development.

Introduction

Single-cell analysis has emerged as a leading methodology for transcriptome analytics. [1] Single-cell data sets (i.e. data involving measurements with single-cell resolution) demonstrate their utility in research contexts for identifying rare subpopulations, characterizing genes that are differentially expressed across conditions, and infering spatio-temporal relationships within the microbiome. [2] Additionally, advances in whole genome amplification and cellular isolation techniques make single-cell data sets more accessible, more informative, and more diverse than ever before. [1] Therefore, there is a clear need to compare, test, and integrate 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 comparing three methods for modeling scRNA-seq expression profiles that account for within-subject correlation differently. We compare theses parameter estimates obtained using data consisting of scRNA-seq observations across multiple subjects with Lupus Neprhitis. General modeling theory is provided in the context of this example and we discuss relevant conclusions, implications, limitations and future research to illustrate our findings.

Previous Results

The following studies use single-cell data to make "down-stream" conclusions. A down-stream analysis will incorperate information generated from a statistical study to make conclusions about relateable biological concepts. During this process, the conclusions drawn from statistical inference are logically equated to biological implications. Therefore, each

"down-stream" result is dependent upon a coherent statistical analysis. The examples below show that coherent statistical inference will be unreasonable when the underlying data exhibits subject-clustering.

Sub-Population Detection

Traditional methods for subpopulation exploration within single-cell data commonly involve unsupervised clustering techniques including Principle Components Analysis (PCA) and K-Nearest Neighbors (KNN). These methods can effectively identify 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 are used to find multi-dimensional boundary values for distinguishing heathly and cancerous bone marrow samples. [4] While all these studies involve single-cell data that incorperates multiple subjects, the modeling methodologies do not provide numerical estimates for the effects of subject-clustered sampling, and therefore can only be used heuristically.

Test for Differential Expression Across Conditions

Single-cell data is used to target treatments by characterizing differential expression across condition. Model-based Analysis of Single-cell Transcriptomics (MAST) is used to compare "primary human non-stimulated" and "cytokine-activated" mucosal-associated invariant T-cells. [5] Additionally, Single-Cell Differential Expression (SCDE) is used to compare 92 embryonic mouse fibroblasts to 92 embryonic human stem cells. [6] Neither of these studies included samples across multiple subjects (excluding paired/treatment sample assumptions used for parameteric tests).

Investigate Spatio-Temporal Microbiome Relationships

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

Description of Motivating Example

Throughout the course of this paper, references are made to "The immune cell landscape in kidneys with lupus nephritis patients" [9]. This paper references single-cell data collected as part of a cross-sectional, case-control study of 27 Lupus Nephritis subjects. Samples of kidney tissue are taken at ten clinical sites across the United States,, where they are crygenically frozen and shiped to a central processing facility. Samples are thawed, dissociated, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytomery [10]. sc-RNA sequencing is performed using a modified CEL-Seq2 method [11], followed by $\sim$ 1 million paired-end reads per cell. Data can be accessed through the ImmPort repository with accession code SDY997.

Data Quality Control

The Seurat Guided Clustering Tutorial [12] is used to examine intial data and perform quality control (QC) filtering. The Seurat package allows for easy classification of low-quality observations by setting threshold values for:

1. the number of unique genes detected in each cell (nFeature), and
2. the percentage of reads that map to the mitochondrial genome (perctMT)

Item (1) is used for identifying empty or broken-cell measurements (indicated by abnormally low gene detection numbers), or duplicate/multiplicate cells measures (indicated by abnormally high gene detection numbers). Item (2) is used to identify dead and/or broken cells since dead or dying cells will retain RNAs in mitochondria, but lose cytoplasmic RNA [2].

The original dristribution of the $PerctMT$ variable across subjexts is displayed in (Figure 1) below:

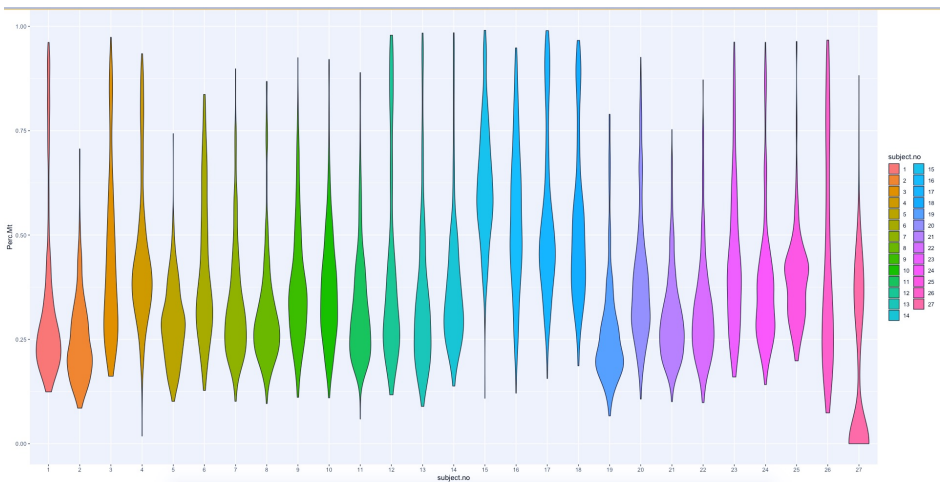

Figure 1:

The QC measures employed by (Arazi A, Rao DA, Berthier CC, et al.) and implemented using the Seurat package required:

1. $1,000 < nFeature < 5,000$
2. $perctMT \leq 25\%$

and the resulting distribution of the *PerctMT* variable is displayed in (Figure 2):

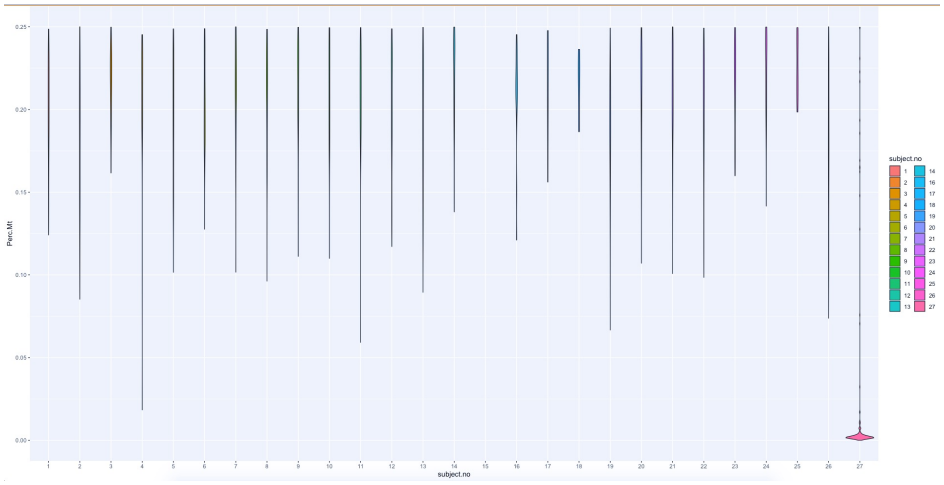

Figure 2:

a decision to increase the *perctMT* threshold to 60% is made to preserve the inherent distribution structure across and within subjects (Figure 3).

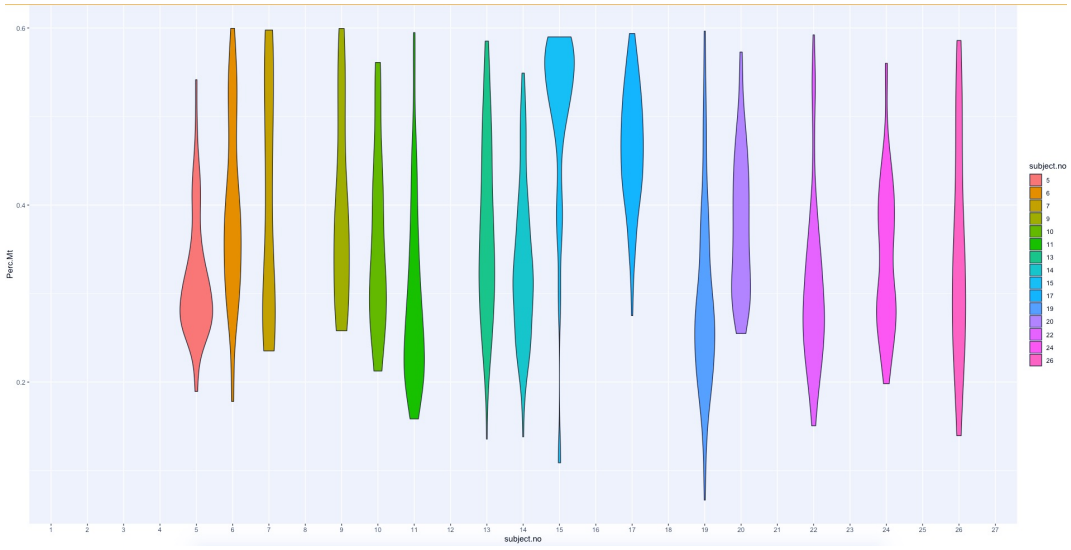

Figure 3:

Further subsetting measures are made to reduce sources of possible conflicting information, by reducing the cellular data types to B-Cells only. This will allow for a more accurate representation of the covariance parameters between-subjects since contributions of variation from inconsistency in cell-structure will be less dramatic.

The data after the updated filters are imposed is summarized in the tables below:

Observation Count Per-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.

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.

9. Arazi A, Rao DA, Berthier CC, et al. (2018) The immune cell landscape in kidneys of lupus nephritis patients. *bioRxiv* 363051.

10. FlowJo X V10. 0.7 r2 flowjo. *LLC https://www flowjo com*.

11. Hashimshony T, Senderovich N, Avital G, et al. (2016) CEL-seq2: Sensitive highly-multiplexed single-cell rna-seq. *Genome biology* 17: 77.

12. Satija R, others (2018) Seurat: Guided clustering tutorial. *Satija Lab http://satijalab org/seurat/pbmc3k_tutorial html*.