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 in data with individual-level clusting. The modeling approaches will be compared theoretically against Linear Regression Models, 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 transcriptome analytics. [1] Single-cell data sets (i.e. data involving measurements with single-cell resolution) have demonstrated 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 have made single-cell data sets more accessible, more informative, and more diverse than ever before. [1]

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 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 all these studies involve single-cell data that 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] Additionally, Single-Cell Differential Expression (SCDE) was used to compare 92 mouse embryonic fibroblasts to 92 embryonic stem cells. [6] Neither of these studies included samples 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 within the microbiome. 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 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.

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 likely 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 suggesting three methods for modeling scRNA-seq expression profiles that account for within-subject correlation differently. We provide a motivating example consisting of scRNA-seq observations across multiple subjects with Lupus Neprhitis. Modeling theory and comparisons will be provided in the context of this

example and the results of the various modeling approaches will be compared. We will discuss relevant conclusions, implications, limitations and future research to illustrate our findings.

Description of Motivating Example

Throughout the course of this paper, references will be made to "The immune cell landscape in kidneys with lupus nephritis patients" [9]. This paper originally references single-cell data collected as part of a cross-sectional, case-control study of 24 Lupus Nephritis (LN) cases and ten control (LD) subjects. Samples of kidney tissue and urine from LN subjects were taken from ten clinical sites across the United States. LD subject samples were obtained at a single site from a living kidney donor, after removal and prior to implantation in the recipient. No LD urine samples were collected. Samples were crygenically frozen and shiped to a central processing facility where they were 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 was performed using a modified CEL-Seq2 method [11], followed by ~ 1 million paired-end reads per cell. Data can be accessed through the ImmPort repository with accession code SDY997.

The Seurat Guided Clustering Tutorial [12] was 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) can be useful for identifying empty or broken-cell measurements (indicated by
87 abnormally low gene detection numbers) and duplicate/multiplicate cells measures (indicated by abnormally high gene detection numbers). Item (2) can help to identify dead and/or broken cells since these cells will retain RNAs in mitochondria, but lose cytoplasmic
90 RNA [2]. Original QC measures employed within the Seurat Package by (Arazi A, Rao DA, Berthier CC, et al.) required:

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

However, after inspecting the impact that the QC filters made on the $perctMT$ distribution (Figure 1)

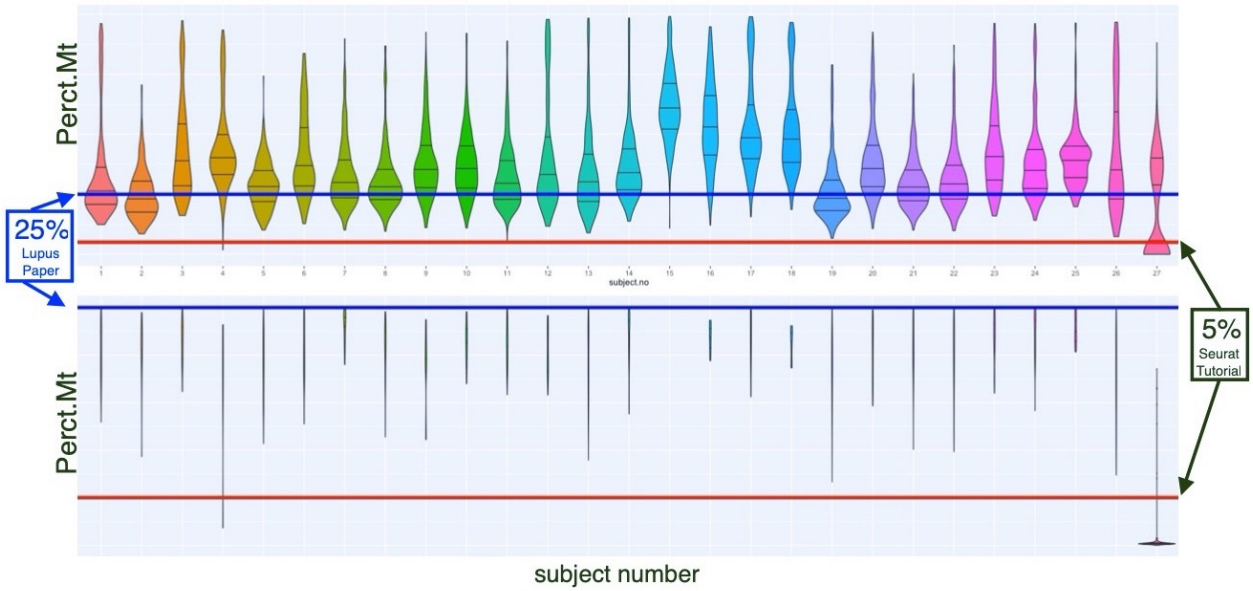
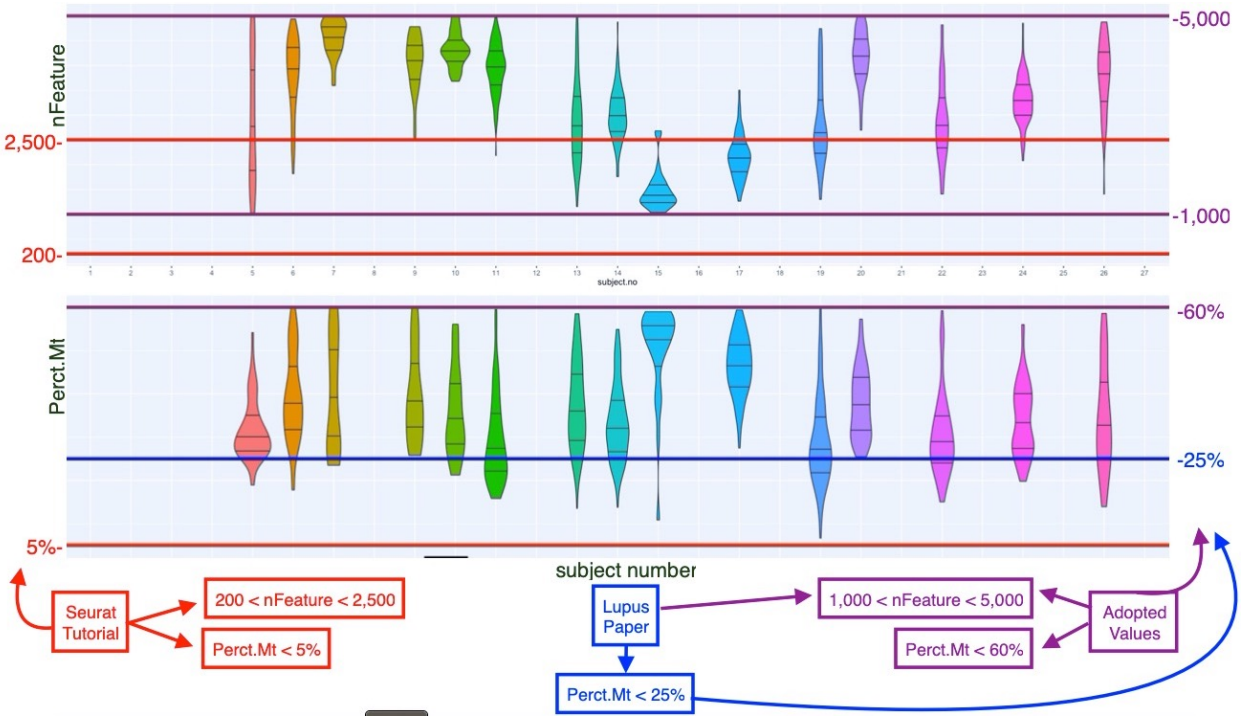


Figure: Perct.mt (pre/post QC filter) by subject with filter values

96

a decision to increase the $perctMT$ threshold to 60% was made to preserve the inherent distribution structure across subjects (Figure 2)



After application of quality control filters, we are left with 1,110 scRNA-seq observations of 38,354 genetic variables distributed across 15 subjects (originally 27 subjects in data, 12 removed due to poor quality data). There are a minimum of 21 and a maximum of 127 observations per subject, with a mean of 74 and a median of 79 observations per subject.

In order to simplify the analysis and make more significant insights into model comparisons, we have decided to choose two pairs of variables from the 38,354 genetic markers to model in a predictor-response relationship. The variables chosen indicated higher values of correlation than arbitrary pairings, and could be associate with important outcomes of interest (e.g. cancer treatment research in the case of MALAT1 [13], or observed limb malformations in the case of FBLN1 variation [14]). An attempt was also made to choose predictor-pairings of interest. The CD19 marker (paired with MALAT1) is a transmembrane protein, encoded by a the CD19 gene. Since the FlowJo cytometry

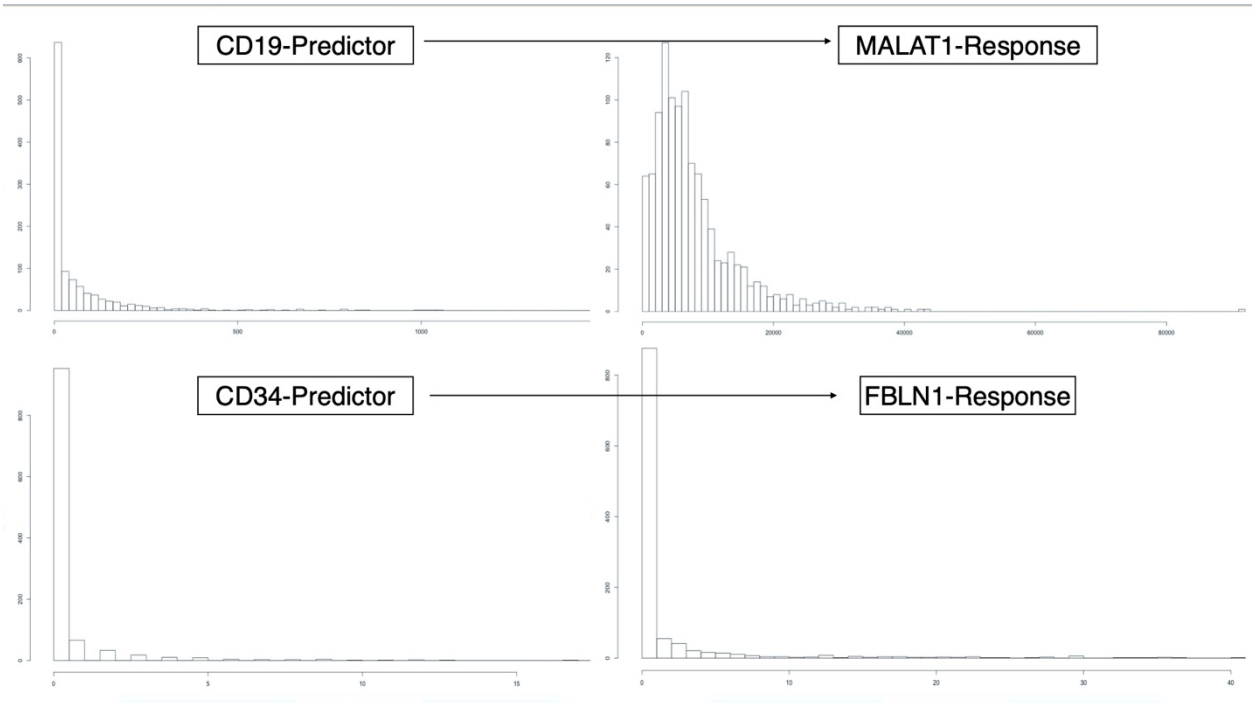
measurement contain CD19 readings, the relationship between a proteomic and transcriptomic predictor and an outcome of interest could be modeled simmultanneously.

114 CD34, the predictor which we will link with FBLN1 is also a transmembrane protein encoded by a gene, and similarly interesting.

Single-cell RNA sequencing data is represented as non-negative integer count data. The

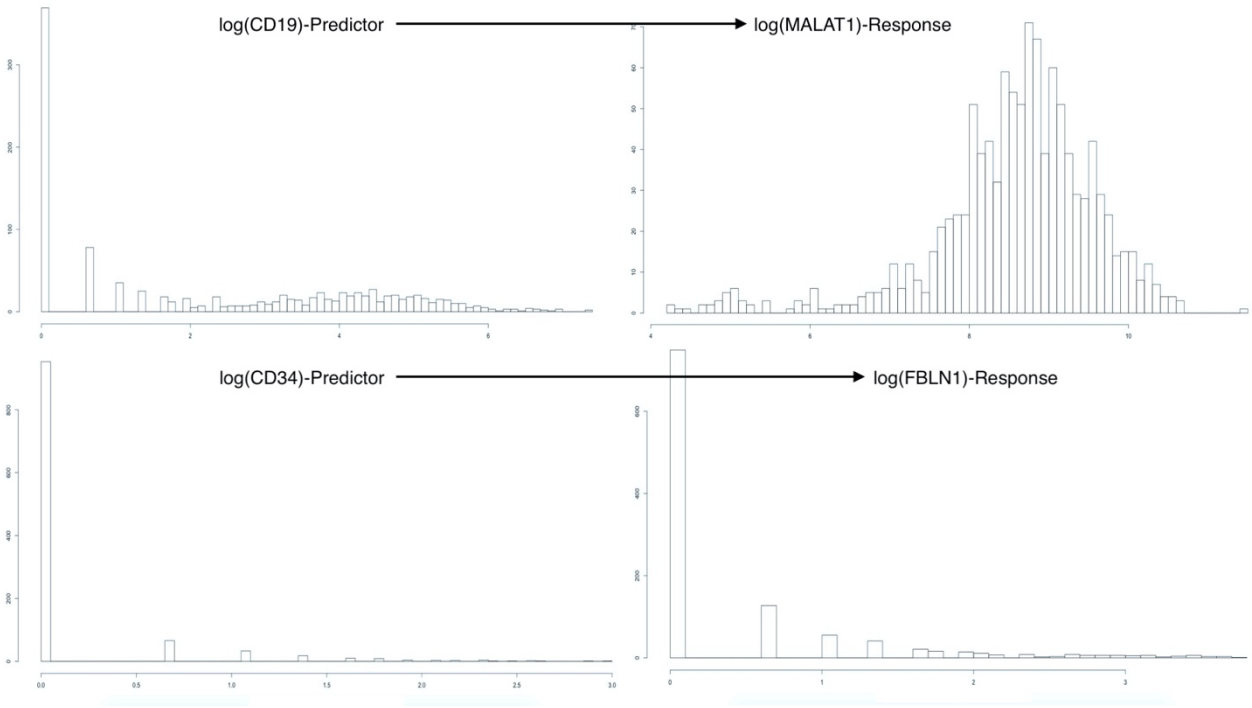
117 observations are specific to much finer genomes (i.e. single-cell resolution), so while the overall scope of gene expression is the same as a traditional bulk experiment, individual observations have a biologically inflated zero-component. Additionally, there are technical

120 zero-inflation components that are associated with protocol variations. Together, these factors contribute to heavily right-skewed variable distributions (Figure 3)



We therefore perform a log-tranform on the predictors and response variables, in an attempt to achieve approximate normality. Figure 4 displays the log transformations given by:

$$X \mapsto \log(X + 1)$$



We can see that the log-transformed response for the MALAT1 ~ CD19 pairing has resulted in an acceptably normal distribution for the response MALAT1. Since the methods we will employ in this paper are only concerned with residual error normality it suffices to have approximately normal response variables (i.e. predictor variable distributions may be reasonably non-normal provided that appropriate residual analysis is performed).

Conversely, the log-transformed response for the FBLN1 ~ CD34 pairing is not inherently better than the un-transformed response. We can clearly see the heavy influence of zero-

inflation in these variables. Even though this is an alarming result, we will continue to model the FBLN ~ CD34 relationship.

Model Parameters

The data we acquired from (Arazi A, Rao DA, Berthier CC, et al) allows us to fully define variables and parameters and outline each model clearly.

We define our outcome(s) of interest to be one of the following transformed variables:

$$R_h = \log(Y_h + 1) \quad \text{for} \quad h = 1,2$$

where

$$Y_1 = \text{MALAT1} \quad \text{and} \quad Y_2 = \text{FBLN1}$$

We aslo define the predictor attached to R_h

$$P_h = \log(X_h + 1) \quad \text{for} \quad h = 1,2$$

where

$$X_1 = \text{CD19} \quad \text{and} \quad Y_2 = \text{CD34}$$

Let a single response be designated as: R_{hij}. The index $i = 1, \dots, 15$ represents the subject from which the observation originated, and the index $j = 1, \dots, n_i$ represents the repeated observation number within subject-i. We note that $n_i \in \{21,22,23, \dots, 127\}$ in the context of the Lupus Nephritis Data. Additionally, the models presented theoretically here (with exception to Model 0) are "Less Than Full Rank" (LTFR) representations of the models for which results will be presented. The Full Rank model results presented create full-rank

model & design matrices by dropping the first level in all factors, and using this as the referrence level.

Linear Regression

We begin the process of comparing models for scRNA-seq expression profiles in Lupus Nephritis subject-clustered tissue data, by describing three linear regression models, with parameters estimated using Least Squares optimizations.

It should be noted that these methods make the assumption that observations are independent, but account for some observational correlation with the use of subject specific intercept and slope terms.

Ultimately, all these methods assume an identical error structure across all observations of the form:

$$\epsilon_{hij} \sim N(0, \sigma_\epsilon^2 * I_{1110})$$

where we are assuming that σ^2 is a fixed variance parameter for all subjects and I_{1110} is the 1110 X 1110 identity matrix.

Simple Linear Regression (Model 0)

Using the notation we defined above, we can write the first model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + \epsilon_{hij}$$

which is equivalent to:

$$\log(Y_{hij}) = \beta_0 + \beta_1 \log(X_{hij}) + \epsilon_{hij}$$

Fixed-Effect Subject-Intercept (Model 1)

Adding a subject-specific intercept term, allows us to account for within-subject correlation by uniformly shifting the fitted values specific to a subject. This model may be written as:

$$R_{hij} = \beta_0 + \beta_{1i}(subject_i) + \beta_2 P_{hij} + \epsilon_{hij}$$

where the term:

$$\beta_{1i}(subject_i) = \begin{cases} \beta_{1i} & \text{if} \quad subject_i = i \\ 0 & \text{if} \quad subject_i \neq i \end{cases}$$

Fixed-Effect Subject-Slope (Model 2)

We may further account for within-subject correlation by adding a term which will ensure that individual subjects' relationships with the covariate of interest is accounted for. This will help to reduce within-subject variation across the predictor space, and will be more noticable for stronger, subject-specific interactions with covariates.

This model may be written as:

$$R_{hij} = \beta_0 + \beta_{1i}(subject_i) + \left[\beta_{2i}(subject_i) * P_{hij}\right] + \beta_3 P_{hij} + \epsilon_{hij}$$

where we are using the same definitions of $(subject_i)$, R_{hij}, and P_{hij} as in Models 0 and 1.

Motivated Results-Linear Regression

The Linear Regression models described in the previous section(s) were each fit in R to the data from the motivating example (Arazi A, Rao DA, Berthier CC, et al) [9]. Estimates for the main intercept (β_0), and the main-effect slope (β_1 in Model 0, β_2 in Model 1, β_3 in Model 2) along with the estimated standard errors, test-statistics, and p-values are displayed in

192 (table 1) and (table 2) for the MALAT1 ~ CD19 relationship and in (table 3) and (table 4) for the FBLN1 ~ CD34 relationship.

Intercept Estimates (MALAT1 ~ CD19)

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 0	8.4618	4.568e-2	1.8526e+2	< 2e-16
Model 1	7.5486	1.2261e-1	6.1564e+1	< 2e-16
Model 2	6.9793	1.3896e-1	5.0226e+1	< 2e-16

Table 1

195

Main-Effect Slope (MALAT1 ~ CD19)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	4.918e-2	1.455e-2	3.381	7.47e-4
Model 1	4.833e-2	1.381e-2	3.500	4.84e-4
Model 2	5.143e-1	6.017e-2	8.546	< 2e-16

Table 2

198

Intercept Estimates (FBLN1 ~ CD34)

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 0	3.510e-1	2.45e-2	1.43e+1	< 2e-16
Model 1	2.7572	6.52e-2	4.23e+1	< 2e-16
Model 2	2.7973	7.68e-2	3.643e+1	< 2e-16

Table 3

201

Main-effect Slope Estimates (FBLN1 ~ CD34)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	7.884e-1	4.92e-2	1.6e+1	< 2e-16
Model 1	1.306e-1	3.42e-2	3.82	1.4e-4
Model 2	8.38e-2	5.89e-2	1.42	1.5492e-1

Table 4

204 We note that in each of the relationships being modeled, we see an decrease in the standard error associated with the main-effect slope as we incorperate subject-specific information regarding the intercept (i.e. when we compare model 0 to model 1).

207 Conversely, we see an increase in the standard error associated with the main-effect slope as we incorperate subject-specific information about the slope.

Model diagnostics including plots of model fit, residual vs fitted value plots, and quantile-

210 quantile residual distribution plots are included as part of the appendix material. *A*

discussion of modeling assumptions will also be included ***(wondering if this needs to be addressed)***

Linear Mixed Models

The next category of approaches to modeling scRNA-seq expression profiles in Lupus Nephritis subject-clustered data we will describe is two distinct Linear Mixed Models. These modeling methods account for subject-clustering differently than the previously discussed Linear Regression models. Linear Mixed Models do not neccessarily assume observational independence, and they can even systematically account for correlation among repeated measures within a subject. Additionally, if we can rationally assume that the responses shown in Figure 3 have a multivariate distribution, the model parameters can be easily estimated using Maximum Likelihood Estimation techniques [15].

Linear Mixed Model with Random Intercept (Model 3)

Model 1 accounted for subject-clustering by assuming that observations within a subject were uniformly influenced by the nested nature (observations within subjects) of the sampling method. However, this assumption is not always be reasonable, as we would expect that responses within each subject would exhibit random variation.

A Linear Mixed Effects Model that includes a Random Intercept accounts for observational correlation due to subject-clustering by assuming that observations within a subject are a consequence of the nested nature of the sampling method, and therefore a consequence of an additive (covariate-independent), subject-specific, effect; AND due to subject-specific

random variation in response measurement associated with measurement instatbility for THAT subject.

This model may be written as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}(subject_i) + \epsilon_{hij}$$

where

$$b_{0i} \sim N(0, \sigma_b^2)$$

$$\epsilon_{hij} \sim N\left(0, \sigma_\epsilon^2 I_{n_i}\right)$$

and we assume that b_{0i} and ϵ_{hij} are independent.

We note that both random-components can be assumed to have a mean of zero because such components are inherently deterministic and can be integrated into the β_0 intercept term.

Linear Mixed Effect Model with Random Slope (Model 4)

Model 2 implemented a Fixed Effect slope in an attempt to reconcile the effects of observational clustering inadequately accounted for by the Fixed Effect Intercept in Model 1. However, in light of the information surrounding the development of Model 3, it is incumbent for us to develop an analogous correction for Model 2. Such a correction would allow us to account for observational correlation due to subeject-clustering as sourced from:

- additive, subject-specific effects due to (subject) clustered sampling methods
- subject-specific random variation associated with measurement instability

- covariate-scaling, subject-specific effects
- covariate-scaling, subject-specific random variation associated with measurement instability

We write this model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}(subject_i) + \left[b_{1i}(subject_i)P_{hij}\right] + \epsilon_{hij}$$

where

$$\mathbf{b} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{G})$$

$$G = \begin{bmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

$$\epsilon_{hij} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}\right)$$

Motivated Results-Linear Mixed Models

The tables displayed (5 - 8) are analogous to Tables (1 - 4) for Model 3 and Model 4.

Intercept Estimates (MALAT1 ~ CD19)

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 3	8.4137	1.1825e-1	7.1151e+1	< 2e-16
Model 4	8.3972	1.3957e-1	6.0166e+1	<2e-16

Table 5

Main-Effect Slope (MALAT1 ~ CD19)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 3	4.920e-2	1.374e-2	3.579	3.6e-4
Model 4	5.938e-2	3.538e-2	1.678	1.19e-1

Table 6

Intercept Estimates (FBLN1 ~ CD34)

Model Number	Estimate	Std. Error	t-Stat	p-Value
Model 3	6.53e-1	2.22e-1	2.94	1.1e-2
Model 4	6.491e-1	2.223e-1	2.92	1.1e-2

Table 7

Main-effect Slope Estimates (FBLN1 ~ CD34)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 3	1.35e-1	3.42e-2	3.95	8.4e-5
Model 4	1.705e-1	7.29e-2	2.34	6.7e-2

Table 8

Again, we note that in each of the relationships being modeled, we see an increase in the standard error associated with the main-effect slope as we incorperate subject-specific information about the slope.

Model diagnostics including plots of model fit, residual vs fitted value plots, and quantile - quantile residual distribution plots are included as part of the appendix material. *A discussion of modeling assumptions will also be included* ***(wondering if this needs to be addressed)***

Sections Still to Come…. GEE, Discussion…etc.

Appendix

Linear Regression Model Diagnostic Plots

This section will contain plots of:

- Model vs Original data
- Fitted vs Original data
- Quantile - Quantile distributions of the model residuals.

Linear Mixed Model Diagnostic Plots

This section will contain plots of:

- Model vs Original data
- Fitted vs Original data
- Quantile - Quantile distributions of the model residuals.

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*.

13. Gutschner T, Hämmerle M, Diederichs S (2013) MALAT1—a paradigm for long noncoding rna function in cancer. *Journal of molecular medicine* 91: 791–801.

14. Debeer P, Schoenmakers E, Twal W, et al. (2002) The fibulin-1 gene (fbln1) is disrupted in at (12; 22) associated with a complex type of synpolydactyly. *Journal of medical genetics* 39: 98–104.

15. Fitzmaurice GM, Laird NM, Ware JH (2012) Applied longitudinal analysis, John Wiley & Sons.