Comparing Models of Subject-Clustered Single-Cell Data

Version 6.0

Lee Panter

Abstract

Single-Cell RNA sequencing data represents a revolutionary shift to approaches being used to decode the human transcriptome. Such data are becoming more prevalent, and are gathered on ever-larger samples of individuals, enabling analysis of subject-level relationships. However, it is not always clear how to conduct this subject-level analysis. Current methods often do not account for nested study designs in which samples of hundreds, or thousands of cells are gathered from multiple individuals. Therefore, there is a need to outline, analyze, and compare methods for estimating subject-level relationships in single-cell expression.

Here, we compare three modeling strategies for detecting subject level associations using single-cell RNA sequencing expression: Linear Regression with Fixed Effects, Linear Mixed Effects Models with Random Effects, and Generalized Estimating Equations. We first present each method. We then compare the regression estimates and standard errors for each method using real single-cell data from a Lupus Nephritis study of 27 subjects. We hoped that this paper presents insights into methods to analyze subject level associations from single-cell expression data.

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 cells, 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 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 noteably on multiple individuals.

The utility of single-cell data, and the feasability 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, we compare three methods for modeling scRNA-seq expression profiles that account for within-subject correlation: Linear Regression with Fixed Effects, Linear Mixed Effects Models with Random Effects, and Generalized Estimating Equations. We will present the framework for each method to reflect the fitting of a predictor-response pairing as defined by: two different Linear Regression linear predictors, two different Linear Mixed Effects linear

predictors, and a single GEE linear predictor. We will assess the estimates assigned to each model for the parameter that reflects subject inspecific interaction between predictor and response (main-effect slope). This parameter will be assesed for stability across model, and across predictor-response pairings 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.

Description of Motivating Example

Throughout the course of this paper, references are made to the 2018 manuscript entitled "The immune cell landscape in kidneys with lupus nephritis patients" [3]. In this manuscript Arazi, Rao, Berthier, et al. compared single-cell kidney tissue sample data from 45 Lupus Nephritis subjects vs. 25 population controls [3]. The kidney tissue samples were collected from ten clinical sites across the United States, were cryogenically frozen, and shipped to a central processing facility. At the central processing facility, the tissue samples were then thawed, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytomery [4]. Single-cell RNA sequencing was performed using modified CEL-Seq2 method [5] with ~ 1 million paired-end reads per cell. The original experimental data may be accessed by visiting the Immport repository with accession code SDY997. Immport-SDY997: https://www.immport.org/shared/study/SDY997

Data Quality Control

The Seurat Guided Clustering Tutorial [6] was used to examine and perform quality control (QC) of the initial data.

This process quantifies the quality of each observation in two numerical measures (based upon two calculated variables, $nFeature$ and $PerctMT$, described below). Threshold values

of these variables can then be chosen and used to filter calls not meeting the chosen criteria. The Seurat tutorial provides methods of automated calculation and filtering implemented by Arazi, Rao, Berthier, et al. in [3]. Identical variable calculations, with alternative threshold settings were independently implemented for this study.

The quality control variables are qualitatively defined as:

1. $nFeature$ is the number of unique genes detected to have a non-zero expression in each cell. This is used to identify cells with an abnormally low or high number of expressed genes. Low numbers may result from empty wells (zero content measurements) or broken-cells, while high numbers may result from observations of more than one cell.
2. $PerctMT$ is the percentage of reads that map to the mitochondrial genome. This 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 pre-QC distribution of $PerctMT$ for each subject is displayed in (Figure 1) below:

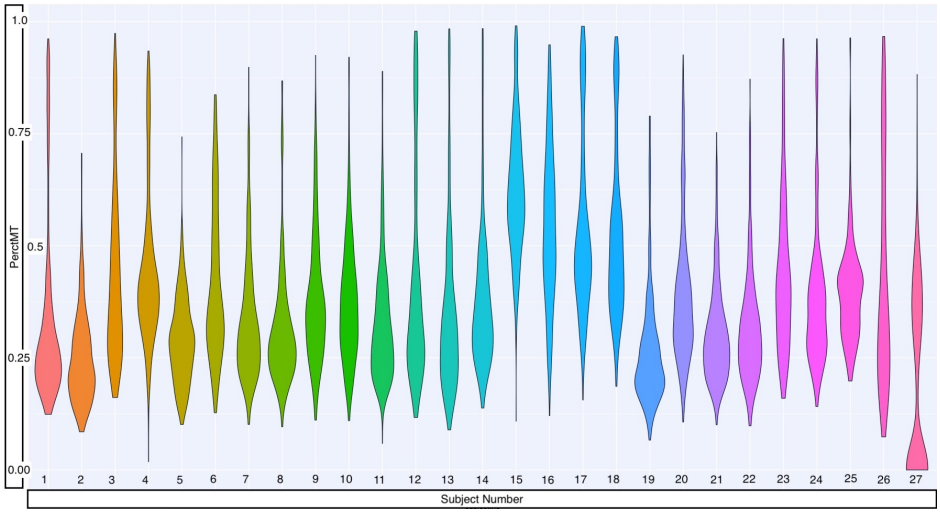

Figure 1: Pre-QC $PerctMT$ Distribution for each subject

The QC measures employed by Arazi, Rao, Berthier, et al. in [3] were:

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

All observations for which the calculated values of $nFeature$ and $PerctMt$ satisfied the inequalites in (1) and (2) above were kept, and the others were considered "low-quality" and removed. The resulting distribution of the $PerctMT$ variable is displayed in (Figure 2):

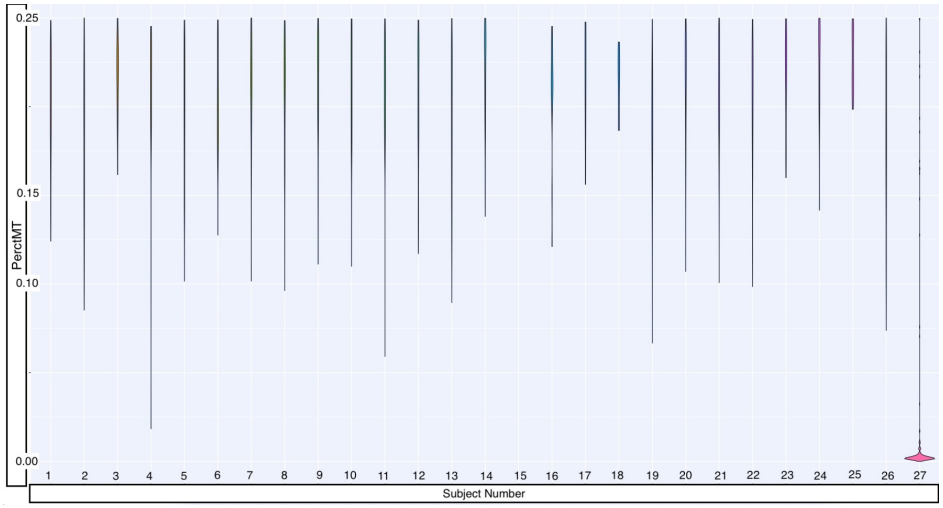

Figure 2: Post QC distribution of $PerctMT$ with thresholds implemented by Arazi, Rao, Berthier, et al

As 84% of cells were removed with the filters chosen by Arazi et al, we chose a more lenient threshold, removing observations with $PerctMT \leq 60\%$ to keep more cells. The additional subsetting measure of restricting the data to only B-cells was made in an effort to regularize (homogenize feature expression) the data sample. The resulting distribution of $PerctMT$ is displayed in (Figure 3) after filtering.

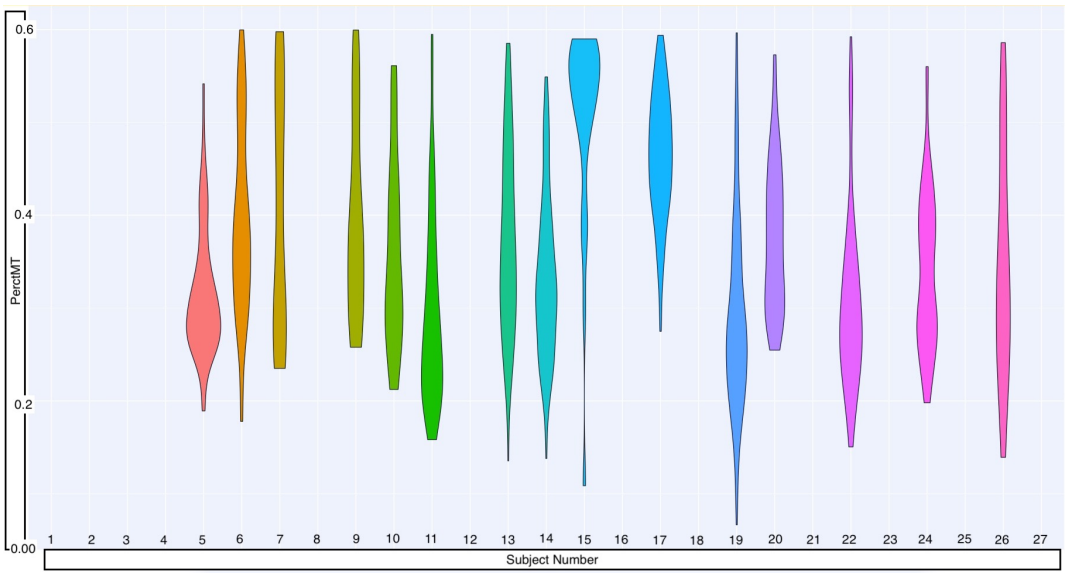

Figure 3: Post QC distribution of $PerctMT$ with thresholds implemented in this paper

The distribution of observations for each subject before and after the quality control thresholds are imposed is also show numerically in Table 1:

Subject Number	1	2	3	4	5	6	7	8	9
Number of Observations Before QC	375	375	364	381	340	383	383	356	372
Number of Observations After QC	0	0	0	0	58	86	32	0	31

Subject Number	10	11	12	13	14	15	16	17	18	19
Number of Observations Before QC	327	311	379	375	345	371	381	381	377	380
Number of Observations After QC	21	107	0	107	100	25	0	122	0	127

Subject Number	20	21	22	23	24	25	26	27
Number of Observations Before QC	381	380	333	333	239	218	378	342
Number of Observations After QC	75	0	87	0	79	0	53	0

Table 1: Observation counts per-subject before and after Quality Control threshold filter restrictions

The process of eliminating observations through quality control threshold measures is compa-

rable to outlier detection and removal. Values defining the quality of an observations are determined by the context of the data being studied, as well as the distribution of values within the data. An observation should only be considered abnormal, poor-quality, uninformative, or unrealistic if it can be characterized as such in the context of its observational setting and compared to the data observed.

The pre-defined thresholds implemented by Arazi, Rao, Berthier, et al outline the expected observational circumstances surrounding the Lupus Nephritis data. However, these limits set unrealistic boundaries in the context of the data provided, and therefore were not reasonable for classifying poor-quality observations.

With this in mind, we also note that quality-control is dissimilar to outlier-detection and removal because the thresholds used define the sample of interest. In this way, an experimenter would conduct quality-control as a sub-sampling method, and would perform outlier detection and removal on the sub-sample.

This subtle, but important difference allows for the *Population of Interest* to be represented by the sample *after QC fiter have been implemented*. This allows us to reduce the data set distribution to subjects with positive observational counts, as they are part of the *Sample of Interest*. This distribution is displayed in Table 2:

Subject Group Number	5	6	7	9	10	11	13	14
Number of Observations	58	86	32	31	21	107	107	100

Subject Group Number	15	17	19	20	22	24	26
Number of Observations	25	122	127	75	87	79	53

Table 2: Observation count per-subject, subjects with positive counts

Table 3 displays the descriptive statistics for the number of observations per-subject.

MIN	1st Q	Median	Mean	3rd Q	MAX
21	42.5	79	74.0	103.5	127

Table 3: observation count per-subject descriptive statistics

Variable Selection and Summaries

We chose two pairs of variables from the 38,354 genetic markers in the Lupus Data to compare across the three methods. The variables we chose have higher values of correlation than arbitrary variable pairings as indicated by a high Pearson Correlation Coefficient (top 10% of all possible pairings), and have previously been associated with human diseases or conditions (e.g. cancer treatment research in the case of MALAT1 [7], or observed limb malformations in the case of FBLN1 [8]). An attempt was also made to assign predictor-pairings of interest. The CD19 marker (paired with MALAT1) is a transmembrane protein, encoded by the CD19 gene. Since the FlowJo cytometry measurements contain CD19 protein readings, the relationship between the "CD19 quantification" used as a predictor predictor and the outcome of interest can be modeled using proteomic or transcriptomics data. CD34, the predictor which we link with FBLN1 is also a transmembrane protein encoded by a gene, and similarly interesting.

Without undergoing the process of expression normalization, single-cell RNA sequencing data is represented as non-negative integer count data. Higher counts correspond to higher detection frequencies and (without compensating for expected expression frequency) these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker.

The variables that we study here are summarized in Appendix Table (A1) - (A4). Each describes selected variable summary statistics (minimum, maximum, average, and median) for the subset samples specific to the subject identifiers used in Table (2).

Measurements of scRNA-seq data can be highly specific to very precise transcriptomic targets (expression profiles can be limited to very small transcriptome scope), so while the agglomerated scope of gene expression across a sample is the same as a traditional bulk experiment, individual observations have a biologically inflated zero-component. There are also *technical* zero-inflation components that are associated with protocol variations, and measurement error.

This is evident in the case of the FBLN1 ~ CD34 pairing, where we see that expression values for several subjects exhibit:

$$\min_j(FBLN1_{ij}) = \min_j(CD34_{ij}) = 0 = \max_j(CD34_{ij}) = \max_j(FBLN1_{ij})$$

where

$$i \in \{5, 6, 7, \ldots, 26\}$$

$$j \in \{1, \ldots, n_i\}$$

Which implies that:

$$(FBLN1_{ij}) = (CD34_{ij}) = 0 = (CD34_{ij}) = (FBLN1_{ij}) \quad \forall i, j$$

We expect the additional presence of zeros to be attributable to both biological and technical sources. Together, these factors contribute to heavily right-skewed variable distributions (Figure 4)

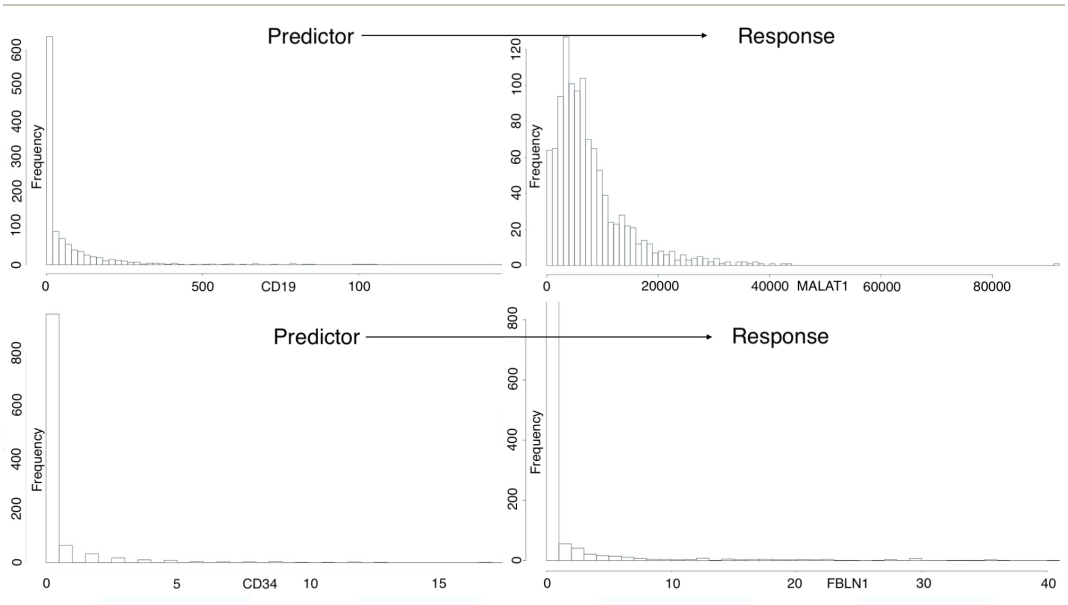

Figure 4: Predictor-Response pairing variable distributions

The MALAT1 variable had a large minimum outcome compared to the other variables. All measurements of this variable are positive in their raw state, so we translate the raw observations negatively by the minimum (67) value. This gives a minimum expression value of zero, which coincides with our intuition as well as the other variables under investigation. It should be noted that this process would be incorperated into the model-fitting procedure automatically through the intercept term.

The modeling methodologies we employ motivates a log-transformation in an attempt to achieve approximate normality, especially for the outcome variable's distribution. We perform the "log plus +1" transformation on all variables:

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

The resulting distributions are shown in Figure (5):

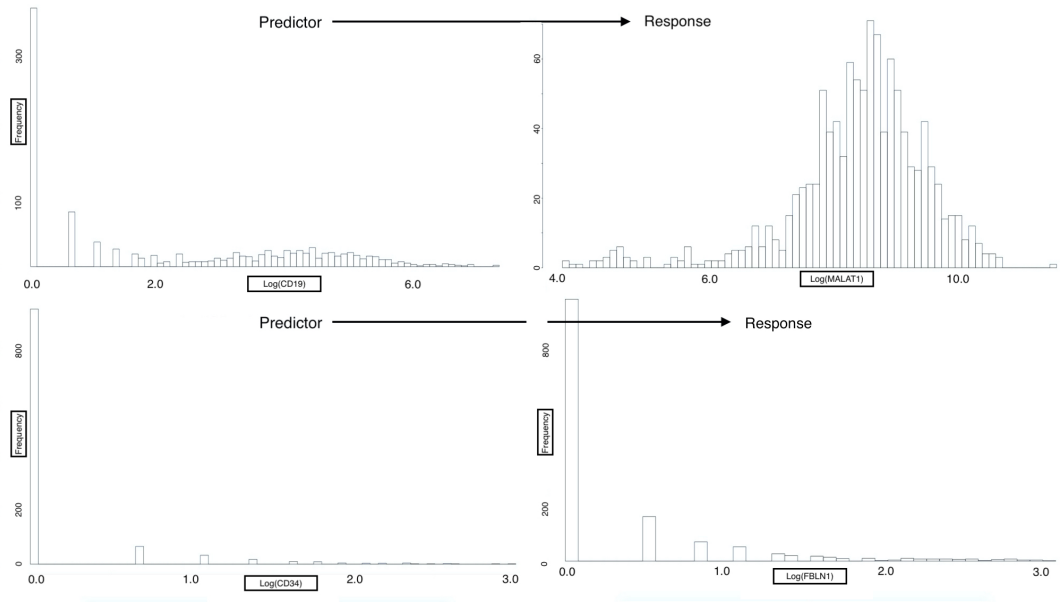

Figure 5: Predictor-Response variable pairings, post-transformation distributions

We see that the log-transformed response MALAT1 is approximately normal distribution. Conversely, the log-transformed response FBLN1 is not inherently better than the untransformed response. We can clearly see the heavy influence of zero-inflation in these variables as is apparent from the dominance of the "zero-bins" in Figure (5).

Regardless, we model each outcome under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. This may not generally the case, and additional transformations or modeling methodologies may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, we will proceed with this assumption and verify ridual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

Model Descriptions

Motivating Example Notation

In the previous sections we described two predictor-response pairings over which the methods described in the sections that follow will be applied. In an effort to provide both a workable, generalized framework, and a clear explanation of the process used to obtain results, an exhaustive explanation of the variable transformation process as applied to the motivating example data is outlined in *Appendix: Derivation of Applied Variables*

In the following sections, we describe each model framework using the generalized predictor-response paring:

$$(X_{ij}, Y_{ij}) \quad \text{for} \quad i = 1, \ldots, N \quad j = 1, \ldots, n_i$$

Where $i = 1, \ldots, N$ represents the observation's *subject of origination* (subject from which the single-cell measurement was taken), and $j = 1, \ldots, n_i$ represents the single-cell measure index taken within subject i (the repeated measure index within each subject).

Linear Modeling

We begin the model framework definitions by describing two linear regression models, with fixed effect parameters estimated using maximum likelihood optimization. It should be noted that linear regression makes the assumption that observations are independent. Linear regression models can account for some structure with the use of a subject specific intercept term as we will see in the second model.

Ultimately, all the methods defined in this section assume an identical error structure across all observations of the form:

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

where we are assuming that σ^2 is the variance parameter for all subjects.

Linear Model (LM)

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

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$$

We note that this model does not account for any structure in the observations, and instead provides estimates for population-averaged relationships:

- The estimated average (across all observations, across all subjects) value of Y_{ij} when $X_{ij} = 0$ (intercept)
- On average (across all observations, across all subjects), the rate of change in Y_{ij} per unit increase in X_{ij} (slope)

Linear Model with Fixed-Effect Intercept (LM-FE)

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

$$Y_{ij} = \beta_0 + \beta_{1i}(subject_i) + \beta_2 X_{ij} + \epsilon_{ij}$$

where we define the term:

$$\beta_{1i}\left(subject_i\right) = \begin{cases} \beta_{1i} & \text{if} \quad subject_i = i \quad \text{for} \quad i = 2, \ldots, N \\ 0 & \text{if} \quad subject_i \neq i \\ 0 & \text{if} \quad i = 1 \end{cases}$$

This model adds $N-1$ estimated parameters $\hat{\beta}_{1i}$ which are the average deviation for each subject from the global estimated mean Linear Model (LM).

Linear Mixed Effects Models

The next category of modeling approaches we describe is linear mixed effect models with random effects. Specifically, we describe two distinct linear mixed effect models that account for subject-correlation in a different manner than the previously discussed linear regression models. Linear mixed efffects models do not require the assumption of independent observations. Correlation structures such as autoregressive, moving-average, or simply unrestricted (unstructured) can be used. Additionally, if we can assume that the model responses have a multivariate normal distribution, the model parameters can be easily estimated using maximum likelihood estimation techniques such as Restricted Maximum Likelihood estimation (REML) [9].

Linear Mixed Effects Model with Random Intercept (LMM-RI)

A random intercept linear mixed effects model (LMM-RI) differs from a linear model with subject specific effects in the way that observational correlation is accounted for. We have seen that such correlation has been accounted for in the LM-FE model with specific mean differences by subject. In order for this method to be justified, it must be the case that observations within a subject are uniformly influenced by the nested nature of the sampling

method. This assumptions is not always reasonable, and a method that allows for responses within each subject to vary randomy according to which subject they belong to, would be more appropriate. A linear mixed effects model with a random intercept controls for subject-level correlations through the use of subject-specific variances, and therefore accomplishes this desired trait. The LMM-RI model is written as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}\,(subject_i) + \epsilon_{ij}$$

where

$$b_{0i} \sim N\left(0, \sigma_b^2\right) \quad \epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$$

$$\text{for} \quad i \in \{1, \ldots, N\} \quad \text{and} \quad j \in \{1, \ldots, n_i\}$$

and we assume that b_{0i} and ϵ_{ij} are independent, and both random-components are assumed to have a mean of zero.

Linear Mixed Effect Model with Random Slope (LMM-RS)

A random slope linear mixed effects model differs from each of the previously considered models because it allows for distinct relationships for each subject between the variables of interest. If we compare a random slope term to a subject-specific fixed effect slope term, we see an analog to the comparison drawn in the description of the LMM-RI model. A model with a subject-specific fixed effect slope term accounts for subject-level observational correlation with subject-specific, predictor-response, mean-difference, relationships. However, it is assumed that observations within subject are uniformly influenced by this relationship due to the nested sampling method. This assumption is not always reasonable, and a method that allows for responses to vary randomly across the predictor-response relationship according to which subject the observation belongs to, would be more appropriate. A linear mixed effects model with a random slope controls for subject-level correlations through the use of

subject-specific variances in the relationships between predictor and response, and therefore accomplished this desired trait. The LMM-RS model is written as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}\left(subject_i\right) + \left[b_{1i}\left(subject_i\right) X_{ij}\right] + \epsilon_{ij}$$

where

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

$$G = \begin{bmatrix} \sigma^2_{b_0} & \sigma_{b_{01}} \\ \sigma_{b_{10}} & \sigma^2_{b_1} \end{bmatrix}$$

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

Generalized Estimating Equations (GEE)

GEE estimates are computed by solving the estimating equation(s) (i.e. solve $U(\beta) = 0$ for β)

$$0 = U(\beta) = \sum_{i=1}^{N} \left\{ \mathbf{D}_i^T \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mu_i\right) \right\} \tag{1}$$

where:

$$\mu_i = \mu_i(\beta) = E\left[\mathbf{Y}_i\right] = \eta_i$$

represents the relationship between the expected value of the response μ_i (not necessarily assumed to be a distribution) and the linear predictor η_i, e.g. in our case we will be assuming

that:

$$\mu_i(\beta) = \eta_i = X_i\beta$$

and

$$\mathbf{D}_i = \begin{bmatrix} \frac{\partial \mu_{i1}}{\beta_1} & \frac{\partial \mu_{i1}}{\beta_2} & \dots & \frac{\partial \mu_{i1}}{\beta_p} \\ \frac{\partial \mu_{i2}}{\beta_1} & \frac{\partial \mu_{i2}}{\beta_2} & \dots & \frac{\partial \mu_{i2}}{\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{in_i}}{\beta_1} & \frac{\partial \mu_{in_i}}{\beta_2} & \dots & \frac{\partial \mu_{in_i}}{\beta_p} \end{bmatrix}$$

is the first derivative matrix, and

$$\mathbf{V}_{hi} = \mathbf{A}_{hi}^{\frac{1}{2}} Corr(\mathbf{Y_{hi}}) \mathbf{A}_{hi}^{\frac{1}{2}}$$

$$\mathbf{A}_{hi} = \underset{n_i}{diag}\left\{\phi_j(t_{ij})\,\nu(\mu_{hij})\right\}$$

We note that $\phi_j(t_{ij})$ and $\nu(\mu_{hij})$ are hyperparameters defined so that we may know the variance as a function of the mean and a scale parameter, i.e:

$$Var(Y_{hij}) = \phi_j(t_{ij})\,\nu(\mu_{hij})$$

The GEE algorithm is iterative and used the following steps to converge at an estimate:

1. Generalized Linear Modeling methods employing Maximum Likelihood Estimation are used to obtain intial estimates for β
2. Estimates for β used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix ($\mathbf{V}_{hi}$) used to obtain new estimates for β by solving (1)
4. Repeat Steps 2 & 3 until algorithm converges

The GEE algorithm has a quality which makes it very appealing for many applications

with observational clustering. Specifically, the algorithm is robust to misspecification of the observational correlation structure. That is, the estimates $\hat{\beta}_{GEE}$ are consistent with β irrespective of the estimates for within-subject correlation.

The GEE algorithm is also very stable, in-part due to the fact that the effect(s) that it estimates are population-averaged. Each of the previous methods (Model 0 withstanding) had subject-specific interpretations, but the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather the population-average.

According to Fitzmaurice, Laird, and Ware [9] we also need to ensure that any responses modeled in the GEE process are stationary, i.e:

$$E\left[Y_{hij}|\mathbf{X}_{hi}\right] = E\left[Y_{hij}|X_{hi1},\ldots,X_{hin_i}\right] = E\left[Y_{hij}|X_{hij}\right]$$

The scRNA-seq data has been assumed to be independent within-subject, therefore we have:

$$E\left[Y_{hij}|X_{hij}\right] = E\left[Y_{hij}|X_{hij'}\right]$$

$$\forall j \in \{1,\ldots,n_i\} \quad j \neq j'$$

as needed.

The three-part specification of the GEE framework includes:

1. The link function and linear predictor
2. Variance function
3. A working covariance matrix

The link function and linear predictor are chosen so that the resulting model estimates will be comparable to preceeding estimates for intercept and slope. Therefore, we will use the identity link function:

$$g(x) = x$$

in conjunction with the linear predictor:

$$g(\mu_{hij}) = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

which implies we will be assuming the general modeling structure:

$$E\left[Y_{hij}\right] = \mu_{hij} = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

we will assume a variance function of the form:

$$Var\left(Y_{hij}\right) = \phi$$

and we will be using a working covariance matrix structure for repeated measures that corresponds to the assumption of independence of observations within a subject.

$$\left[Corr\left(Y_{hij}, Y_{hik}\right)\right]_{jk} = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$$

$$for \quad j, k \in \{1, \ldots, n_i\}$$

Results

Table 8 and table 9 display parameter value estimates, standard errors, test statistics, and p-values for the main-effect slope term estimated by all five modeling approaches:

Coefficient Estimates

(MALAT1 ~ CD19)

Model Designation	Model Description	Estimate	Std. Error	Test Statistic	p-value
LM	Linear Model	4.918e-2	1.455e-2	3.381*	7.47e-4
LM-FE	Linear Model with Fixed-Effect Intercept	4.833e-2	1.381e-2	3.500*	4.84e-4
LMM-RI	Linear Mixed Model with Random Intercept	4.920e-2	1.374e-2	3.579*	3.6e-4
LMM-RS	Linear Mixed Model with Random Slope	5.938e-2	3.538e-2	1.678*	1.19e-1
GEE	Generalized Estimating Equations	4.918e-2	1.455e-2	3.381**	7.47e-4

Table 8: Summary Table for $CD19 \sim MALAT1$ variable parings. * Approximate normal distribution. ** Approximate Wald-Z distribution

(FBLN1 ~ CD34)

Model Designation	Model Description	Estimate	Std. Error	Test Statistic	p-value
LM	Linear Model	7.884e-1	4.92e-2	4.002*	<2e-16
LM-FE	Linear Model with Fixed-Effect Intercept	1.31e-1	3.42e-2	3.818*	1.42e-4
LMM-RI	Linear Mixed Model with Random Intercept	1.35e-1	3.42e-2	3.95*	8.4e-5
LMM-RS	Linear Mixed Model with Random Slope	1.705e-1	7.29e-2	2.34*	6.7e-2
GEE	Generalized Estimating Equations	7.884e-1	4.92e-2	4.002**	< 2e-16

Table 9: Summary Table for $CD34 \sim FBLN$ variable parings. * Approximate normal distribution. ** Approximate Wald-Z distribution

The main-effect slope parameter represents subject-inspecific information about how predictor and response variables are correlated. We have seen that each method accommodates the effects of subject-level correlation differently. Specifically, we noted that the Linear Model and GEE methods estimated population-averaged parameters, whereas the other models had subject-specific interpretations. So, when the main-effect slope parameter estimate is compared across methods with otherwise identical in structure, we can directly attribute changes to a shift in this parameters value to be a redistribution of the attributed source of correlation between the variables of interest.

The percent change in the main-effect slope parameter across models is displayed for each variable paring in Tables 10 and 11. Values are full-percentage changes, and are calculated using:

$$\text{Percent Change}[A]_{ij} = \left(\frac{A_j - A_i}{A_i}\right) * 100$$

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-1.7283	0.0407	20.7401	0.0000
LM-FE	1.7587	0	1.8001	22.8636	1.7587
LMM-RI	-0.0407	-1.7683	0	20.6911	-0.0407
LMM-RS	-17.1775	-18.6090	-17.1438	0	-17.1775
GEE	0.0000	-1.7283	0.0407	20.7401	0

Table 10: Main effect slope Percent Change matrix, $CD19 \sim MALAT1$ variable pairing

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-83.3841	-82.8767	-78.3739	0.0000
LM-FE	501.8321	0	3.0534	30.1527	501.8321
LM-RI	484.0000	-2.9630	0	26.2963	484.0000
LM-RS	362.4047	-23.1672	-20.8211	0	362.4047
GEE	0.0000	-83.3841	-82.8767	-78.3739	0

Table 11: Main effect slope Percent Change matrix, $CD34 \sim FBLN$ variable pairing

It is worthwhile to comment on the consistency properties of estimates across models within variable parings. In each of the variable paring scenarios we see that changes between models within either of the cases:

1. LM $\Leftrightarrow$ GEE (identical estimates/zero percent change)
2. LM-FE $\Leftrightarrow$ LMM-RI $\Leftrightarrow$ LMM-RS

results in smaller percent-change values than changes between the cases. Changes between LMM-RS and LM-FE/LMM-RI in the CD19-MALAT1 variable paring are technically higher, but this result is most likely an artifact of subject-specific interactions with the covariate. Since models within each of the cases (1) and (2) above estimate similarly interpreted parameters (subject-specific vs population averaged) the result is otherwise expected.

Standard Error Estimates

The standard errors for this parameter are also enlightening when compared across models. A change in a parameter estimate's standard error across modeling methodology represents a revision in the underlying distributional conclusions the method is using to support its result. In this way, an increased standard error between two models that are estimating the same parameter indicates a decrease in obtained (obtainable) estimate precision.

Tables 12 and 13 are percent change matrices for the standard error of the main effect slope parameter:

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-5.0859	-5.5670	143.1615	0.0000
LM-FE	5.3584	0	-0.5069	156.1912	5.3584
LMM-RI	5.8952	0.5095	0	157.4964	5.8952
LMM-RS	-58.8751	-60.9666	-61.1645	0	-58.8751
GEE	0.0000	-5.0859	-5.5670	143.1615	0

Table 12: Main effect slope Standar Error Percent Change matrix, $CD19 \sim MALAT1$ variable pairing

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-30.4878	-30.4878	48.1707	0.0000
LM-FE	43.8596	0	0.0000	113.1579	43.8596
LM-RI	43.8596	0.0000	0	113.1579	43.8596
LM-RS	-32.5103	-53.0864	-53.0864	0	-32.5103
GEE	0.0000	-30.4878	-30.4878	48.1707	0

Table 13: Main effect slope Standar Error Percent Change matrix, $CD34 \sim FBLN$ variable pairing

Changes in standard errors display similarly infomative consistencies. In each variable pairing:

1. The standard error increases on the following model transitions:
    a. All Other Models $\longrightarrow$ LMM-RS
    b. LMM-RI $\longrightarrow$ All Other Models
2. The standard error decreases on the following model transitions:
    a. All Other Models $\longrightarrow$ LMM-RI
    b. LMM-RS $\longrightarrow$ All other Models
    c. LM $\longrightarrow$ LM-FE
    d. GEE $\longrightarrow$ LM-FE

a. The modeling transitions in (1a) correspond with the addition of information to the model in the form of a subject-specific "Random Effect Slope".
b. The transitions in (1b) correspond to either:
    i. addition of the parameter in 1a
    ii. loss of subject-specific information that was originally incorperated into the variance-component of the model. I.e., loss of subject-specific variability information.
c. The transitions in (2a and 2b) are the inverse representation of the relationships outlined in (a) and (b) above.

d. The transition in (2c) corresponds to the incorperation of additive, subject-specific, predictor independent (mean-effect) information into the model.

e. The transition in (2d) corresponds to the addition of the assumption of independence between subjects, along with a purely parametric fitting method (for LM-FE).

The preceding relationships allow us to deduce the effects of the various types of information inclusion on the precision of parameters used to make inferences on the relationship between predictor and response. Beneficial (increases in precision) information inclusions will result in reductions to standard error estimates (section 2 transitions, with explanations c and d above). Detrimental (decreases in precision), or contradictory information will result in increased standard error estimates (section 1 transitions, a & b explanations).

Explanation (e) demonstrates the importance of considering the effect of correlation between subjects in single-cell data. Since this transition is non-zero, it is clear that there is an effect associated with subject-clustered sampling. Otherwise the use of an independence assumption in an analysis would lead to identical results as an analysis without this assumption.

Test Statistics

Tables 14 and 15 are percent change matrices for the test statistic of the main effect slope parameter:

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	3.5197	5.8563	-50.3697	0.0000
LM-FE	-3.4000	0	2.2571	-52.0571	-3.4000
LMM-RI	-5.5323	-2.2073	0	-53.1154	-5.5323
LM-RS	101.4899	108.5816	113.2896	0	101.4899
GEE	0.0000	3.5197	5.8563	-50.3697	0

Table 14: Main effect slope Test Statistic Percent Change matrix, $CD19 \sim MALAT1$ variable pairing

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-4.5977	-1.2994	-41.5292	0.0000
LM-FE	4.8193	0	3.4573	-38.7114	4.8193
LM-RI	1.3165	-3.3418	0	-40.7595	1.3165
LM-RS	71.0256	63.1624	68.8034	0	71.0256
GEE	0.0000	-4.5977	-1.2994	-41.5292	0

Table 15: Main effect slope Test Statistic Percent Change matrix, $CD34 \sim FBLN$ variable pairing

If we look back to Tables 8 and 9, we can see that the sign of the test statistic for each model remains positive across each method, as well as across each variable pairing. The test-statistic distributions, which are approximately normal, or Wald-Z (asymptotically z-distributed) can be used to justify a symmetry argument that larger-valued test statistics represent higher significance of the parameter. This is analogous to higher magnitude, normally distributed, test statistics representing higher significance.

The patterns that we observe in the test statistic percent change matrices serves to largely reinforce previous conclusions we have made using the estimates of coefficients or standard errors. Transitions between LM/GEE and LM/LMM-RI tend to result in larger test statistic changes than changes within models that estimate similar parameters (i.e LM-FE $\leftrightarrow$ LMM-RI which both estimate subject-specific parameters, and LM $\leftrightarrow$ GEE which both estimate population-averaged parameters).

Transitions to LMM-RS from any model results in a loss of significance of the main effect slope parameter as indicated by negative test statistic percent changes. This aligns with our intuition as we would expect the average relationship between outcome and covariate to diminish as emphasis on subject-specific relationship between outcome and covariate increased.

The results outlined in the section above are all based on the inclusion of various types of subject-specific information. These relationships can be classified according to how they affect our ability to perform inference on the relationship between a predictor and a response using subject-correlated scRNA-seq data. To this effect, we can now evaluate our variable-pairing relationship(s) to determine if there is a significant effect from the nested sampling methods used to create the scRNA-seq data,

and if there is an effect, how can this effect best be accounted for.

Discussion

We have compared three methods of modeling scRNA-seq data, each accounting for subject-level associations in a different manner. We analyzed two different Linear Models, a population-average Ordinary Least Squares model, and a Linear Model with a subject-specific Fixed Effect. Our second method included two different types of Linear Mixed Effects Models. We fit a Random Intercept Model, and a Random Slope Model. Finally, we fit another population-average model using the Generalized Estimating Equations algorithm.

The primary goal of our analysis has been to address the arising presence of scRNA-seq data sets gathered on larger samples of individuals, and specifically the lack of clarity surrounding methods to conduct subject-level analyses using them. In order to achieve this goal, we described the consistency of estimates across modeling methodologies for a parameter intended to appraise the population-averaged relationship between two scRNA-seq variables. This approach allows us to examine the magnitude, direction, and significance of subject-correlation as it is included in a variety of methods.

Our results indicated that methods evaluating similarly interpreted parameters (i.e. population-averaged vs subject-specific) had more similar (or identical) parameter estimate outcomes than the dissimilarly interpreted modeling approaches. We also noticed a consistent increase in parameter standard error upon the inclusion of a random slope.

Even though such patterns may be diagnosable with just two scRNA-seq variable pairings, more would be needed to make significant conclusions regarding further parameter stability trends. The evaluation of more variable pairings is the foremost objective left outstanding in this analysis. Supplementary variable pairings would serve to reinforce current findings and stabilize estimate trends heavily related to subject-specific features.

Although the Seurat Guided Clustering Tutorial [6] provides a framework for quality control with

integrated exploratory analysis, the observed protocol dependencies of scRNA-seq data must still be considered before any analysis can be conducted. While methods of combining existing scRNA-seq data have been used to successfully integrate multiple-subjects' single-cell observations [10], no batch-effect corrections or expression normalization has been performed to account for sources of possible confounded or misrepresented subject-level correlation effects.

As single-cell RNA sequencing data sets rise in pervasiveness, the need for subject-level analysis in data sets that are subject-correlated will also rise. This paper presented a foundational comparison for such an analysis. It is hoped that this paper has presented unique insights into the methods and analyses of subject-level associations in scRNA-seq data.

Appendix

Derivation of Applied Variables

Tables

Table A1

CD19 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	0	678	36.6724	0.0
6	0	299	36.6860	7.5
7	0	10	2.1250	1.0
9	0	1052	89.4194	3.0
10	0	158	37.5714	2.0
11	0	339	28.3178	1.0
13	0	629	56.0841	18.0
14	0	251	42.2600	19.0
15	0	148	26.6000	0.0
17	0	982	112.3770	16.0
19	0	665	59.3386	5.0
20	0	287	40.1200	23.0
22	0	380	43.4483	1.0
24	0	282	55.0127	27.0
26	0	1624	268.4151	110.0

Table A1: Predictor $CD19$ variable summaries ($CD19 \sim MALAT1$)

Table A2

MALAT1 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	67	40812	10206.3621	9195.0
6	757	30774	11568.2791	11689.0
7	441	17916	6868	4039.5
9	311	18239	5703.9355	5983.0
10	1875	17160	6638.5714	6190.0
11	349	34082	9716.0280	8826.0
13	99	25572	5867.9439	4895.0
14	355	15740	6154.1500	5720.5
15	157	11923	3839.0800	3467.0
17	337	8342	2960.2541	2692.0
19	227	91961	13959.9843	10125.0
20	379	21736	7301.4133	6417.0
22	161	28429	6881.7471	5068.0
24	240	42792	6248.8228	5955.0
26	1114	32426	8463.1698	6426.0

Table A2: Response $MALAT1$ variable summaries ($CD19 \sim MALAT1$)

Table A3

CD34 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	0	19	3.0517	1
6	0	0	0	0
7	0	0	2	1
9	0	6	0.4516	0
10	0	5	0.6667	0
11	0	7	1.2056	1
13	0	0	0	0
14	0	1	0.4000	0
15	0	0	0	0
17	0	0	0	0
19	0	0	0	0
20	0	2	0.1867	0
22	0	4	0.3563	0
24	0	5	0.2911	0
26	0	0	0	0

Table A3: Predictor $CD34$ variable summaries ($CD34 \sim FBLN1$)

Table A4

FBLN1 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	3	41	19.3448	18
6	0	0	0	0
7	0	16	4.2500	3
9	0	8	1.8710	1
10	0	30	11.9524	10
11	0	8	1.5140	1
13	0	1	0.0093	0
14	0	5	0.5700	0
15	0	1	0.0400	0
17	0	3	0.0246	0
19	0	2	0.0157	0
20	0	9	2.5867	2
22	0	11	0.9885	0
24	0	4	0.4557	0
26	0	0	0	0

Table A4: Response $FBLN1$ variable summaries ($CD34 \sim FBLN1$)

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

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

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

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

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

8. 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.

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