Comparing Models of Subject-Clustered Single-Cell Data

Lee Panter

*Audrey Hendricks, PhD-Committee Chair and Advisor**

Stephanie Santorico, PhD-Committee Member†

Rhonda Bacher, PhD-Committee Member‡

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 RNA sequencing expression.

Here, I compare five modeling strategies for detecting subject level associations using single-cell RNA sequencing expression: linear regression, linear regression with subjects modeled as fixed effects, linear mixed effects models with subjects modeled as random intercepts and random slopes, and generalized estimating equations. I first present each method. I then compare the regression estimates and standard errors for each method using real single-cell data from a Lupus Nephritis study of 27 subjects. I hope that this paper presents insights into methods to analyze subject level associations from single-cell expression data.

---

*The University of Colorado-Denver
†The University of Colorado-Denver
‡The University of Florida

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 cell sample 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 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 notably on multiple individuals.

The utility of single-cell data, and the feasibility 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, I compare five methods for modeling scRNA-seq expression profiles that account for within-subject correlation: linear modeling (LM), linear modeling with subjects as fixed effects (LM-FE), linear mixed effects models with subjects as random intercepts (LMM-RI) and random slopes (LMM-RS), and generalized estimating equations. I will present the framework for each method in the context of a generalized predictor-response pairing. Then I will

assess each model's estimate of the fixed effect slope parameter for stability across modeling approach 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 this paper references are made to the 2018 article entitled "The immune cell landscape in kidneys with lupus nephritis patients", in which Arazi, Rao, Berthier, et al. compare 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, cryogenically frozen, then 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 cytometry [4]. Single-cell RNA sequencing was performed using a 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

I use the Seurat Guided Clustering Tutorial [6] to perform quality control (QC) of the initial data. This process quantifies the quality of each single-cell observation in two numerical measures (based upon two calculated variables, **nFeature** and **PerctMT**). Threshold values of these variables are chosen and used to filter cells (observations) 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 are independently implemented for this study.

The quality control variables are conceptually 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 (partial) 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 as 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:
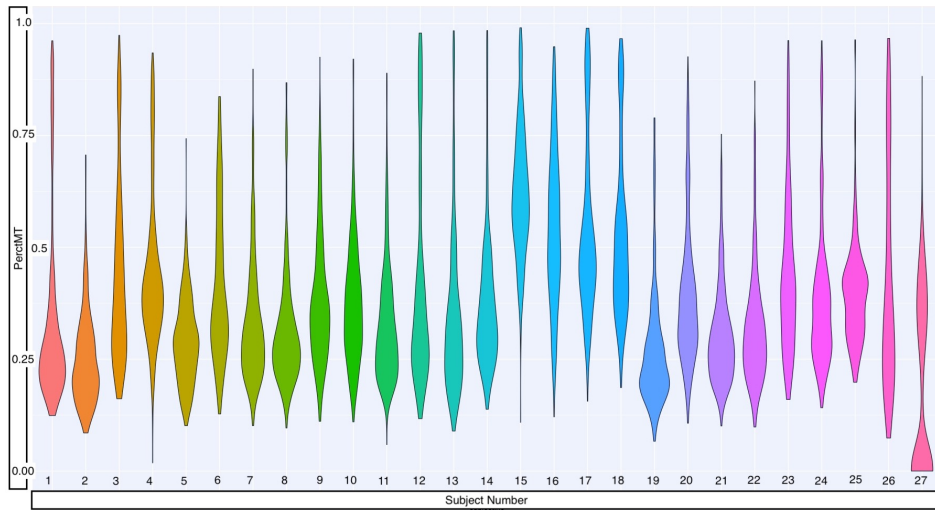

Figure 1: Pre-QC **PerctMT** Distribution for each subject

The QC measure thresholds employed by Arazi, Rao, Berthier, et al. in [3] are:

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

All observations for which the calculated values of **nFeature** and **PerctMT** satisfied the inequalities 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**):

4
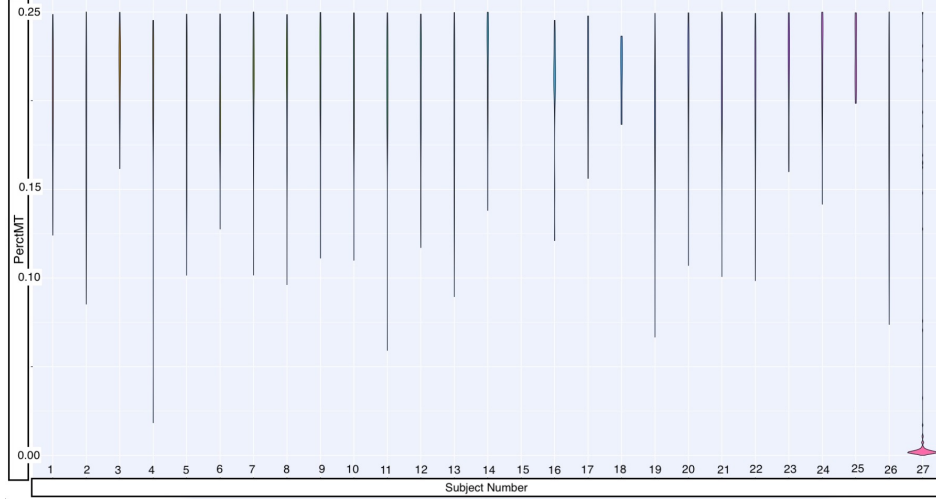
Figure 2: Post QC distribution of **PerctMT** with thresholds implemented by Arazi, Rao, Berthier, et al

As 84% of cells as removed with the filters chosen by Arazi et al, I choose a more lenient threshold, removing observations with **PerctMT** $\leq 60\%$, in an effort to keep more cells.

An additional restriction of the data to only B-cells is made in an effort to regularize the data sample (i.e. homogenize feature expression). The resulting distribution of **PerctMT** is displayed in (**Figure 3**) after filtering.
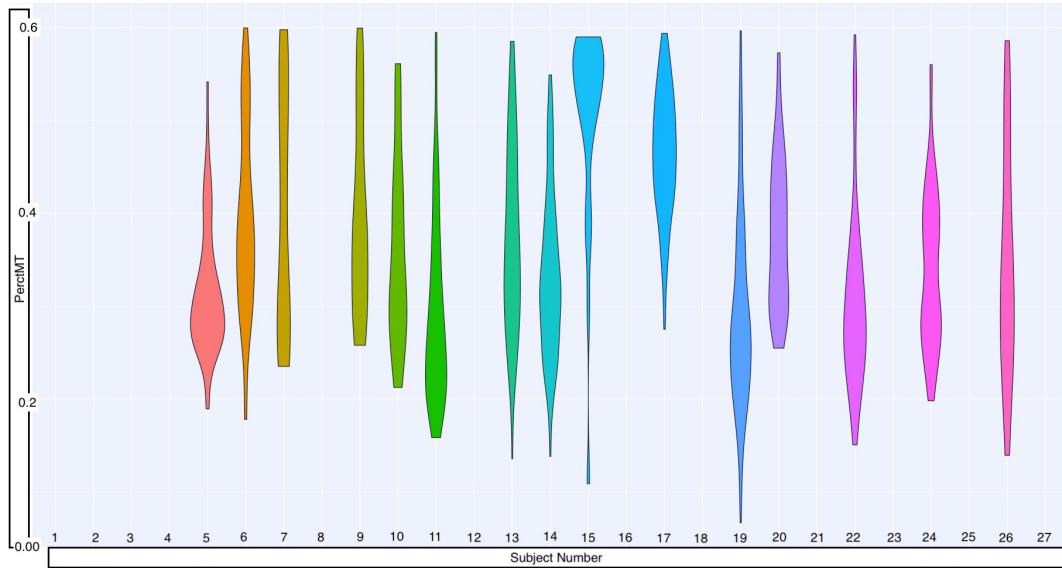
Figure 3: Post QC distribution of **PerctMT** with thresholds implemented in this paper

The distribution of observations for each subject Pre/Post QC (with updated **PerctMT** $\leq 60\%$ threshold value) is shown numerically in (**Table 1**):

Subject Number	1	2	3	4	5	6	7	8	9
NO of Obs Before QC	375	375	364	381	340	383	383	356	372
NO of Obs After QC	0	0	0	0	58	86	32	0	31
Subject Number	10	11	12	13	14	15	16	17	18
NO of Obs Before QC	327	311	379	375	345	371	381	381	377
NO of Obs After QC	21	107	0	107	100	25	0	122	0
Subject Number	19	20	21	22	23	24	25	26	27
NO of Obs Before QC	380	381	380	333	333	239	218	378	342
NO of Obs After QC	127	75	0	87	0	79	0	53	0

Table 1: Observation counts per-subject Pre/Post QC with updated **PerctMT** $\leq 60\%$ threshold value.

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

Table 2: observation count per-subject summary statistics Post QC with updated **PerctMT** $\leq 60\%$ threshold value

Variable Selection and Summaries

I select two pairs of variables from the 38,354 genetic markers in the Lupus Data to compare across the five modeling methods. The variables I choose have higher values of correlation than arbitrary variable pairings as indicated by a high Pearson Correlation Coefficient (both selected pairings are within the top 10% of highest Pearson Correlation Coefficients 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]-used as the first outcome, or

observed limb malformations in the case of FBLN1 [8]-used as the second outcome). I also attempt to assign predictor-pairings of interest. The CD19 marker (the predictor paired with MALAT1) is a transmembrane protein encoded by the CD19 gene. The FlowJo cytometry measurements contain CD19 protein readings, so the relationship between CD19 as a predictor and the outcome of interest (MALAT1) 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 values. Higher counts correspond to higher detection frequencies and these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker (e.g CD19, CD34, MALAT1, FBLN1).

I provide numerical summaries of the four selected variables in Appendix **Tables (A1) - (A4)**. Each describes selected variable summary statistics (minimum, maximum, average, and median) for the positive observational count subjects in (**Table 1**).

Measurements of scRNA-seq data are specific to precise transcriptomic targets. This means that single-cell expression profiles (a single observation) can be limited to a small transcriptomic scope. So while the agglomerated scope of gene expression across a sample is the same as (or broader than) 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. Together, these factors contribute to right-skewed variable distributions (**Figure 4**)
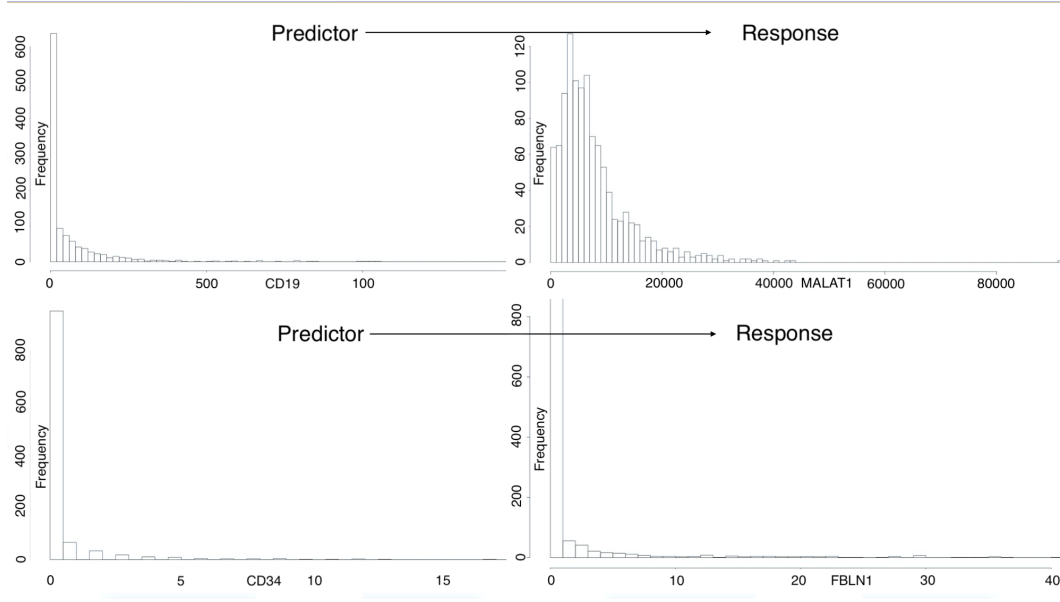
Figure 4: Predictor-Response pairing variable distributions

The MALAT1 variable has a large minimum outcome compared to the other variables, so I translate all the values of this variable *negatively* by the minimum value.

$$min(\text{MALAT1}) = 67$$

This gives a minimum expression value of zero, which coincides with intuition as well as the minimum value of the other variables under investigation.

The modeling methodologies I employ motivate a log-transformation in an attempt to achieve approximate normality, especially for the outcome variable's distribution. I perform the "log plus +1" transformation on all variables (predictor and outcome):

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

The resulting distributions are shown in (**Figure 5**)
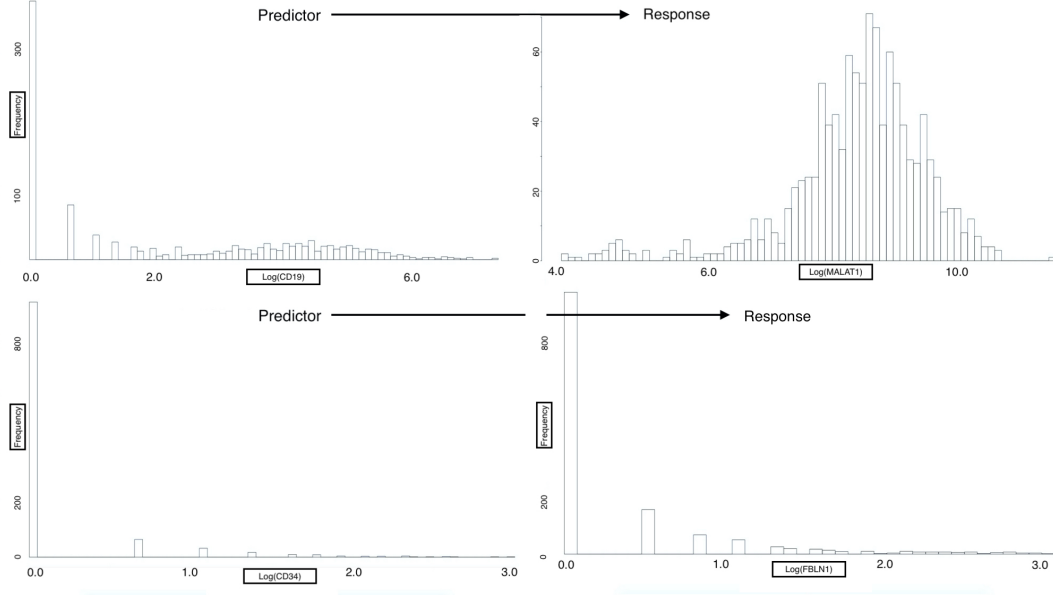
Figure 5: Predictor-Response variable pairings, post-transformation distributions

The log-transformed response of MALAT1 is approximately normally distributed; however, the log-transformed response FBLN1 is not inherently better than the un-transformed response.

Regardless, each outcome is modeled under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. It may be the case that additional transformations and/or alternative modeling techniques may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, I proceed with this assumption and I verify residual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

Model Descriptions

In the following sections a description is provided for each modeling approach using a framework that is generalizable to arbitrary predictor-response pairings of the form:

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

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

Linear Model (LM)

The linear model may be written as:

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

It should be noted that this model does not account for correlation structure in the data, and instead assumes observational independence. Additionally, this method generates estimates of parameters that have population average interpretations.

The error term in the model, represented by ϵ_{ij} is assumed to be a normally distributed random variable with mean zero and variance σ_ϵ^2.

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

Adding a subject-specific intercept term to the LM model allows for the accounting of *some* 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

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

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

Linear Mixed Effects Models

Linear mixed effects models that incorporate subjects using random effects are the next methods outlined. Linear mixed effects 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 it can be reasonably assumed 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. Observational 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 randomly 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:

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

where

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

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

it is assumed that b_{0i} and ϵ_{ij} are independent.

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

A random slope linear mixed effects model differs from each of the previously considered methods because it allows for distinct relationships for each subject between the predictor and response variables of interest. A model with a subject-specific <u>fixed effect</u> slope term accounts for subject-level observational correlation with expected differences between the subject-specific, predictor-response relationships. However, this method still assumes that observations within subjects are uniformly influenced as a result of this due to the nested sampling method. Sometimes, a method that allows for responses to vary randomly across the predictor-response relationship according to which subject they belong 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

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

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

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

Generalized Estimating Equations (GEE)

The Generalized Estimating Equation (GEE) method focuses on estimating the fixed effect parameters specified within a model. The method does not directly estimate variances and covariances, but instead approximates these values using a "working variance-covariance" structure.

Procedurally GEE estimates are computed by finding numerical solutions for an *estimating equation* $U(\beta) = 0$, where

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

and

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

$$= X_i^T\beta$$

outlines the dependence of the marginal expectation of the response on the covariates through the link function $g()$, and therefore:

$$\mu_i(\beta) = g^{-1}\left(\eta_i\right) = g^{-1}\left(E\left[\mathbf{Y}_i|X_i\right]\right)$$

$$= g^{-1}\left(X_i^T\beta\right)$$

The first derivative matrix $\mathbf{D_i}$ is given by:

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

$\mathbf{V}_i$ is the working covariance matrix. This matrix can be estimated from hyper parameters created in the process of numerically solving for the GEE estimates β_{GEE} that solve $U\left(\beta\right) = 0$. Please see Fitzmaurice, Laird, and Ware [9] for more information.

The GEE algorithm uses the following general 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 (from 1) used to compute hyper-parameters

3. New estimates for hyper-parameters and working covariance matrix ($\mathbf{V}_i$) used to obtain new estimates for β by solving (1)

4. Repeat Steps 2 & 3 until algorithm converges

The algorithm is robust to misspecification of the observational covariance structure. So initially incorrect specifications of the working covariance matrix still converge to the appropriate structure form with algorithmic iteration.

The GEE algorithm is stable, in-part due to the fact that the method estimates population-average effects. Each of the previous methods (model LM withstanding) have subject-specific interpretations, but the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather a *hypothetical* average subject.

According to Fitzmaurice, Laird, and Ware [9], responses modeled with the GEE process need to be stationary, i.e:

$$E\left[Y_{ij}|\mathbf{X}_i\right] = E\left[Y_{ij}|X_{i1}, \ldots, X_{in_i}\right] = E\left[Y_{ij}|X_{ij}\right]$$

The scRNA-seq data is assumed to be independent within-subject, therefore:

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

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

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

All three items are chosen so that the resulting model estimates are comparable to preceding

estimates for intercept and slope. Therefore, the identity link function will be assumed:

$$g(x) = x$$

in conjunction with the linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_i$$

which implies the modeling equation:

$$E[Y_{ij}] = \mu_{ij} = \eta_{ij} = \beta_0 + \beta_1 X_{ij}$$

Additionally, an identity variance function will be assumed:

$$Var(Y_{ij}) = \phi\nu(\mu_{ij})$$
$$= \phi\mu_{ij}$$

and a working covariance matrix structure for repeated measures that correspond to the assumption of independence of observations within a subject will be specified as:

$$[Cov(Y_{ij}, Y_{ik})]_{jk} = \begin{cases} Var(Y_{ij}) & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$$

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

Parameter Interpretations

The GEE and LM modeling techniques are methods of obtaining estimates of population-averaged parameters. These parameter values are interpreted as contributing to the response

of the *hypothetical* average subject (not representative of any single subject within the sample).

Suppose that $\hat{\beta}_{population}$ represents an estimate obtained for the fixed effect slope as obtained by one of the previously describe *population-averaged* modeling methods. An interpretation of this parameter is: **across all subjects, a one-unit increase in the predictor (X_{ij}) is associated with a $(\hat{\beta}_{population})$ unit change in the expected outcome (Y_{ij})**

Conversely, the LM-FE, LMM-RI and LMM-RS modeling techniques are methods of obtaining estimates of subject-specific parameters. These parameter values are interpreted as contributing to the average response having controlled for a constant subject of origin (i.e. the parameter estimate attributable to a single subject within the sample)

Suppose that $\hat{\beta}_{subject}$ represents an estimate obtained for the fixed effect slope as obtained by one of the previously describe *subject-specific* modeling methods. An interpretation of this parameter is: **for a given subject (controlling for the subject of origin) a one-unit increase in the predictor (X_{ij}) is associated with a $(\hat{\beta}_{subject})$ unit change in the expected outcome (Y_{ij}).**

Results

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 3. $MALAT1 \sim CD19$ **model estimates:** Fixed effect slope estimate, standard error, test statistic, and p-value for each model for the relationship between the predictor $CD19$ and the outcome $MALAT1$. ** Approximate Wald-Z distribution.

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 4. $FBLN1 \sim CD34$ **model estimates:** Fixed effect slope estimate, standard error, test statistic, and p-value for each model for the relationship between the predictor $CD34$ and the outcome $FBLN1$. ** Approximate Wald-Z distribution.

Displayed in **tables (5)** - **(10)** below are percent change in: parameter estimate, standard error, and test statistic for the $MALAT1 \sim CD19$ variable pairing in **tables (5)-(7)** and the $FBLN1 \sim CD34$ variable paring in **tables (8)-(10)**. Where the percent change in estimate is defined as:

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

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 5: Main effect slope percent change matrix, $MALAT1 \sim CD19$ variable pairing

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 6: Main effect slope standard error percent change matrix, $MALAT1 \sim CD19$ variable pairing

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 7: Main effect slope test statistic percent change matrix, $MALAT1 \sim CD19$ 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 8: Main effect slope percent change matrix, $FBLN1 \sim CD34$ 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 9: Main effect slope standard error percent change matrix, $FBLN1 \sim CD34$ 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 10: Main effect slope test statistic percent change matrix, $FBLN1 \sim CD34$ variable pairing

Parameter Value Comparisons

A comparison of main effect slope coefficient, standard error and test statistic across modeling approaches within variable pairings indicates that estimates produced by the LM and GEE methods are numerically indistinguishable at resolutions greater than 10^{-4}. The LM-FE and LMM-RI method estimates are also similar since estimates for each parameter type (coefficient, standard error and test statistic) exhibit magnitude and directional similarities in both variable pairings.

The LMM-RS estimates for the fixed effect slope parameter standard error is the highest when compared to the corresponding estimates within variable pairing as generated by other modeling methods. In contrast, the standard error of the fixed effect slope parameter is lowest for the LMM-RI model within variable pairings. The LM-FE model has a lower fixed effect slope standard error than either the LM or the GEE model within both variable combinations.

The differences in test statistics of the fixed effect slope parameter for each modeling method within each variable pairing are analogous to the differences in slope coefficients previously noted. In particular, test statistics have similar values between the LM and GEE models as well as between the LM-FE and LMM-RI models. Test statistics calculated for the LMM-RS model have the most irregular values, and also result in calculated p-values that have decreased significance in up to three orders of magnitude of percent change.

Nested Model Comparisons

Variable Pair	Model	Resid DF	RSS	DF	Sum of Squares	F-stat	P(>F)
MALAT1-CD19	LM	1108	1167.76				
	LM-FE	1094	935.89	14	231.87	19.36	6.4776e-44
FBLN1-CD34	LM	1108	650.51				
	LM-FE	1094	214.92	14	435.59	158.38	2.8058e-251

Table 11: ANOVA nested model comparison table for testing the inclusion of the subject-specific fixed-effect intercept

(**Table 11**) above is a nested model comparison, the result of which is an F-test statistic indicating that there is very strong evidence to support the inclusion of the subject-specific fixed-effect intercept into the LM model.

Variable Pair	Model	df	AIC	logLik	L.Ratio	p-value
MALAT1-CD19	LM	3	3224.097	-1609.048		
	LMM-RI	4	3032.024	-1512.012	194.0722	4.1068e-44
FBLN1-CD34	LM	3	2572.807	-1283.403		
	LMM-RI	4	1438.086	-715.043	1136.72	3.4517e-249

Table 12: ANOVA nested model comparison table for testing the inclusion of the subject-specific random effect intercept

(**Table 12**) above is a nested model comparison, the result of which is a likelihood ratio statistic indicating that there is very strong evidence to support the inclusion of the subject-specific random effect intercept into the LM model.

Variable Pair	Model	df	AIC	logLik	L.Ratio	p-value
MALAT1-CD19	LMM-RI	4	3032.024	-1512.012		
	LMM-RS	6	2993.820	-1490.910	42.20503	6.8437e-10
FBLN1-CD34	LMM-RI	4	1438.086	-715.043		
	LMM-RS	6	1438.068	-713.034	4.018095	0.1341

Table 13: ANOVA nested model comparison table for testing the inclusion of the subject-specific random effect slope

(**Table 13**) above is a nested model comparison, the result of which is a likelihood ratio statistic idicating that there is very strong evidence to support the inclusion of the subject-specific random effect slope into the LMM-RI model for the $MALAT1 \sim CD19$ variable pairing. However, there is insufficient evidence to support the inclusion of the subject-specific random effect slope into the LMM-RI model for the $FBLN1 \sim CD34$ variable pairing.

Discussion

Here, I compare five modeling strategies for detecting subject level associations in single-cell RNA sequencing data gathered over 27 subjects from a Lupus Nephritis study. I compare estimates of a fixed effect slope parameter generated by five modeling techniques: linear modeling (LM), linear modeling with subjects modeled as fixed effects (LM-FE), linear mixed effects models with subjects modeled as random intercepts (LMM-RI) and random slopes (LMM-RS), and generalized estimating equations (GEE).

I find that population average models (i.e. LM and GEE) and subject specific intercept models (i.e. LMM-RI and LMM-RS) tend to produce similar results within the same model class (population average or subject specific intercept) but different results between model classes. The highest standard errors are indicated in the LMM-RS model, and the lowest standard errors in the LMM-RI model. LM-FE standard error is also found to be smaller than both LM and GEE standard error values. Nested model comparisons indicate that inclusion of subject specific terms is advisable

at all levels (fixed and random, intercept and slope) with exception of the random slope in the $FBLN1 \sim CD34$ variable paring.

Interpretations of subject specific parameters are contextually authentic provided that they are used in inference conditional to their subject of origin. Conversely, interpretations of population average parameters are accurate when they are used for inference on a population's hypothetical representation of centrality. Under conditions of linearity and error normality, it can be shown that subject specific parameters are marginal representations of population average parameters. This distinction explains the parameter estimate disparities as estimated between the LM/GEE methods compared to the LM-FE/LMM-RI methods.

This analysis is subject to several drawbacks and limitations. All the results are based on evidence obtained from just two single-cell RNA sequencing variable pairings. In the future, comparing the consistency of these models over all model pairs is needed. Additionally, single-cell RNA sequencing data is heavily influenced by protocol dependencies and measurement inconsistencies. Quality control must be carefully considered and conducted prior to any analysis.

The utility and promise of single-cell RNA sequencing data indicates that such data will become more prevalent and will be extended to multiple subject samples. I have presented an initial comparison of methods for detecting subject-level associations in single-cell RNA sequencing data sets.

Appendix

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.