Comparing Models of Subject-Clustered Single-Cell Data

Version 4.0

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:

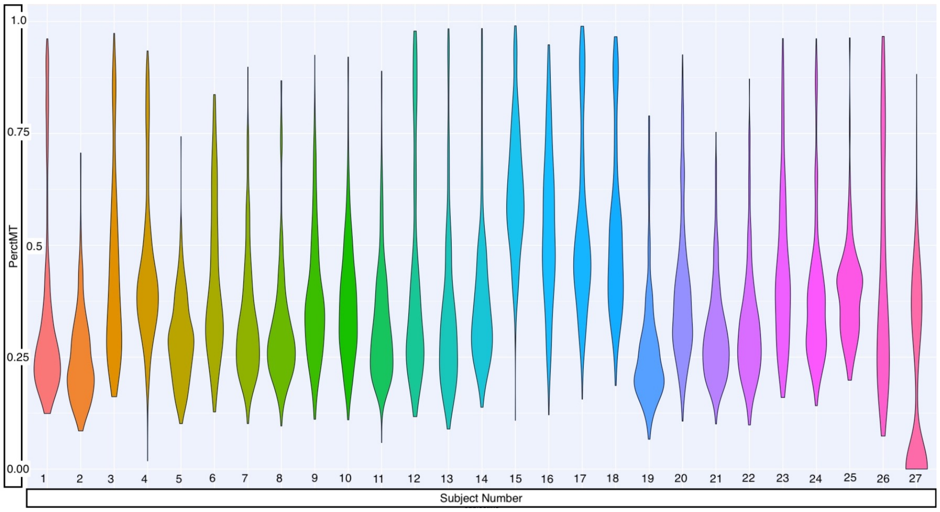

Figure 1: Pre-QC $PerctMT$ Distribution for each subject

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):

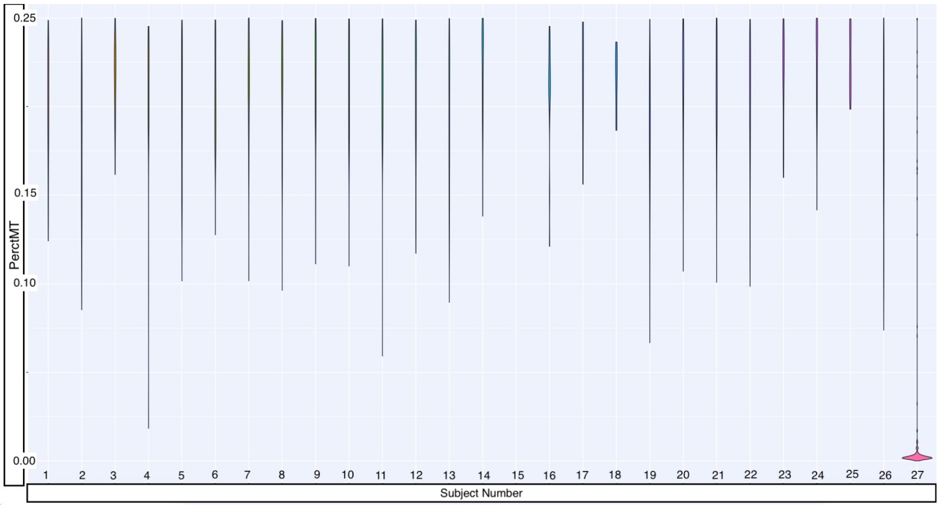

Figure 2: Post QC distribution of $PerctMT$ with thresholds implemented by Arazi, Rao, Berthier, et al

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.

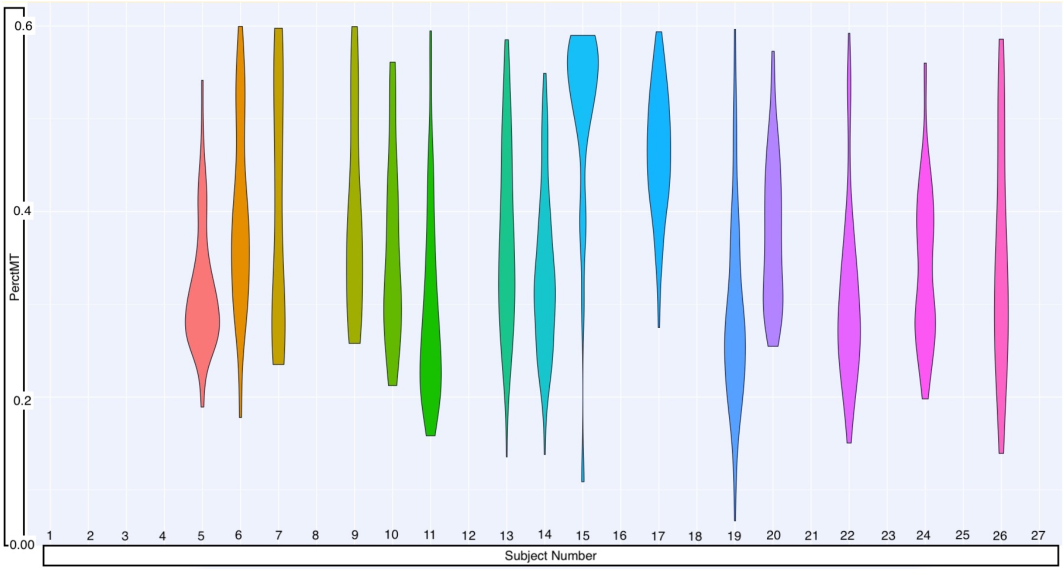

Figure 3: Post QC distribution of $PerctMT$ with thresholds implemented in this paper

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 comparable 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	15	17	19	20	22	24	26
Number of Observations	58	86	32	31	21	107	107	100	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:

$$\underset{j}{min}(FBLN1_{ij}) = \underset{j}{min}(CD34_{ij}) = 0 = \underset{j}{max}(CD34_{ij}) = \underset{j}{max}(FBLN1_{ij})$$

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)

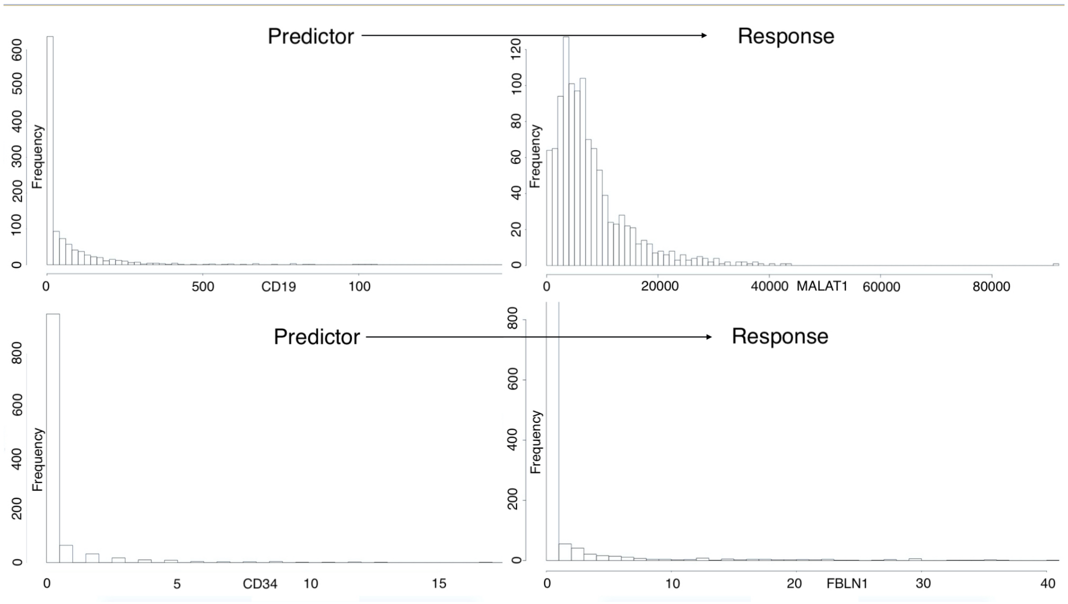

Figure 4:

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:

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

The resulting distributions are shown in Figure (5):

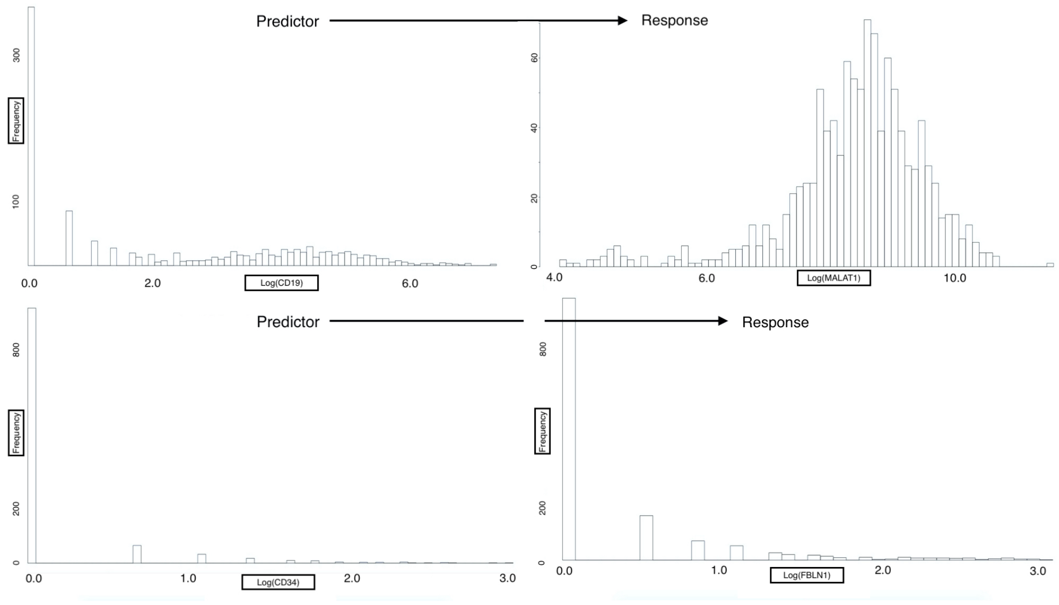

Figure 5:

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

We define our outcome(s) of interest to be one of the following transformed variables as taken from Arazi, Rao, Berthier, et al. Let a single observation be designated as: R_{hij}. The index $h = \{1, 2\}$ represents the model pairing number ($CD19 \sim MALAT1$ is pairing #1 $CD34 \sim FBLN1$ is pairing #2) $i \in \{5, 6, \ldots, 26\}$ represents the subject (name of subject by number) from which the observation originated, and the index $j = 1, \ldots, n_i$ represents the single-cell observation within subject-i. We note that $n_i \in \{21, 22, 23, \ldots, 127\}$ in the context of the Lupus Data.

We perform the transformations:

$$R_h \;=\; \log\left(Y_h^* + 1\right)$$

where

$$Y_h^* = Y_h - min(Y_h)$$

giving

$$Y_1^* = \text{MALAT1} - 67 \quad \text{and} \quad Y_2^* = \text{FBLN1}$$

We aslo define the predictor attached to R_h as:

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

where

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

We present the theoretical model frameworks here as “Less Than Full Rank” (LTFR) representations. The Full-Rank model results presented in the *Results* section to follow are created by droping the first level in all factors and using this as the referrence level.

Linear Regression

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 these methods make the assumption that observations are independent, and should therefore be used for comparison to modeling methods to come. However, the linear regression models we present here can account for some observational correlation 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:

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

where we are assuming that σ^2 is a common 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 write the first model as:

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

which is equivalent to:

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

We note that this model does not account for observational correlation, and instead provides an estimation for population-averaged relationships, namely:

- What is the estimated average (across all observations, across all subjects) value of R_{hij} when $P_{hij} = 0$
- On average (across all observations, across all subjects) what is the average rate of change in R_{hij} per unit increase in P_{hij}

Fixed-Effect Subject-Specific 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 we define the term:

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

This model provide the added estimated parameter $\hat{\beta}_{1i}$ which tells us a uniform estimated average deviation for each subjects' fitted response from the global estimated mean provided by Model 0 (Simple Linear Regression).

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 neccessarily assume independence of observations. Correlation structures such as AR(1), spatial power, or unstructured can be used to estimate parameters determining correlation amongst observations within a subject and between observations (across subjects). Additionally, if we can rationally assume that the responses shown in Figure 3 have a multivariate normal distribution, the model parameters can be easily estimated using Maximum Likelihood Estimation techniques [9].

Linear Mixed Effects Model with Random Intercept (Model 2)

Model 1 (Linear Regression with Fixed Effect Intercept) accounts for subject correlation by assuming that observations within a subject are uniformly influenced by the nested nature of the sampling method (i.e. observations are sampled so that they are identically correlate within each subject). However, this assumption may not always be reasonable, as we could imagine that responses within each subject also exhibit random variation that is related to nested sampling methods.

A Linear Mixed Effects Model that includes a Random Intercept accounts for subject-level observational correlation by inducing individual-specific levels of random variation into all observations specific to each subject. Such a model may be written as:

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

where

$$b_{0i} \sim N\left(0, \sigma_b^2\right) \quad \text{for} \quad i \in \{5, 6, \ldots, 26\}$$

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

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 as non-zero components are inherently deterministic and can be integrated into intercept terms.

Linear Mixed Effect Model with Random Slope (Model 3)

A further accounting for the effects of subject-level observational correlation may be made by extrapolating on Model 2 (Linear Mixed Effects Model with Random Intercept) with the addition of a random intercept.

The incoperation of a Fixed Effect subject-specific slope would account for subject-level observational correlation by assuming that the relationship between predictor and response are uniformly influenced across observations. Implying that, in addition to the average response devation from the estimated average response, there is also an average uniform shift in how each subjects' response changes with respect to a unit shift in the predictor. Again, this assumption may not be reasonable, as we may expect variation in how responses within a subject deviate from the estimated average change in response over the predictor space.

We will therefore incorperated a Random Slope into the format of the Random Intercept model (Model 2) to attempt to reconcile these effects. This will allow for us to account for observational correlation due to subject-level sampling as sourced from:

- subject-specific random variation associated with measurement instability
- predictor-dependent, subject-specific random variation associated with measurement instability

We write this model as:

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

where

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

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

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

Generalized Estimating Equations (Model 4)

Our final method for modeling scRNA-seq expression profiles is Generalized Estimating Equations (GEE). Dissimilar to each of the methods previously described, GEE regression esitimates are obtained using methodologies that allow for non-continuous responses. GEE also extrapolates on the techniques used for modleing non-normal responses by incorperating the effects of observational correlation.

GEE estimates are computed by solving the estimating equation(s):

$$0 = U(\beta) = \sum_{i=1}^{15} \left\{ \mathbf{D}_{hi}^T \mathbf{V}_{hi}^{-1} \left(\mathbf{y}_{hi} - \mu_{hi} \right) \right\} \tag{1}$$

where:

$$\mu_{hi} = \mu_{hi}(\beta) = E\left[\mathbf{Y}_{hi}\right] = \eta_{hi}$$

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,

$$\mathbf{D}_{hi} = \begin{bmatrix} \frac{\partial \mu_{hi1}}{\beta_1} & \frac{\partial \mu_{hi1}}{\beta_2} & \cdots & \frac{\partial \mu_{hi1}}{\beta_p} \\ \frac{\partial \mu_{hi2}}{\beta_1} & \frac{\partial \mu_{hi2}}{\beta_2} & \cdots & \frac{\partial \mu_{hi2}}{\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{hin_i}}{\beta_1} & \frac{\partial \mu_{hin_i}}{\beta_2} & \cdots & \frac{\partial \mu_{hin_i}}{\beta_p} \end{bmatrix}$$

is the first derivative matrix, and

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

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

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

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

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.

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

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

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:

(MALAT1 ~ CD19)

Model Description	Estimate	Std. Error	t-Stat	p-value
Ordinary Least Squares Model 0	4.918e-2	1.455e-2	3.381	7.47e-4
Linear Regression Fixed Effect Intercept Model 1	4.833e-2	1.381e-2	3.500	4.84e-4
Linear Mixed Model Random Intercept Model 2	4.920e-2	1.374e-2	3.579	3.6e-4
Linear Mixed Model Random Slope Model 3	5.938e-2	3.538e-2	1.678	1.19e-1
Generalized Estimating Equations Model 4	2.22e-1	6.3e-1	1.24**	2.2e-1

Table 8: Summary Table for $CD19 \sim MALAT1$ variable parings. (**Approximate t-distribution using chi-square distributed Wald test statistic)

(FBLN1 ~ CD34)

Model Description	Estimate	Std. Error	t-Stat	p-value
Ordinary Least Squares Model 0	7.884e-1	4.92e-2	1.6e+1	< 2e-16
Linear Regression Fixed Effect Intercept Model 1	1.306e-1	3.42e-2	3.82	1.4e-4
Linear Mixed Model Random Intercept Model 2	1.35e-1	3.42e-2	3.95	8.4e-5
Linear Mixed Model Random Slope Model 3	1.705e-1	7.29e-2	2.34	6.7e-2
Generalized Estimating Equations Model 4	7.88e-1	2.2e-1	1.281e+1**	3.4e-4

Table 9: Summary Table for $CD34 \sim FBLN$ variable parings. (**Approximate t-distribution using chi-square distributed Wald test statistic)

The main-effect slope parameter represents subject-inspecific information about how the predictor and response variables are correlated. The modeling approaches being compared here all estimate this effect; however, each method accomodates the effects of subject-level correlation differently. When the main-effect slope parameter estimate is compared across methods, we can directly attribute changes in the parameter's value to a shift in attributed association between subject-specific and population-averaged effects.

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:

$$\text{Percent Change}[A]_{ij} = \frac{A_i - A_j}{\left(\frac{A_i + A_j}{2}\right)}$$

Model	Mod 0	Mod 1	Mod 2	Mod 3	Mod 4
Mod 0	0	-1.74	0.04	18.8	127.0
Mod 1	1.74	0	1.78	20.5	128.0
Mod 2	-0.04	-1.78	0	18.8	127.0
Mod 3	-18.8	-20.52	-18.75	0	116.0
Mod 4	-127.46	-128.49	-127.43	-115.6	0

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

Model	Mod 0	Mod 1	Mod 2	Mod 3	Mod 4
Mod 0	0	-143.16	-141.52	-128.9	-0.05
Mod 1	143.16	0	3.31	26.5	143.13
Mod 2	141.52	-3.31	0	23.2	141.50
Mod 3	128.9	-26.5	-23.2	0	128.85
Mod 4	0.05	-143.13	-141.50	-128.85	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 amongts the models in the set $S_1 = \{1, 2, 3\}$ result in smaller changes in the main effect slope coefficient than changes beteween the models in set S_1 and those models in $S_2 = \{0, 4\}$. Sincle the models in S_1 provide estimates of individual-level association, but the models in S_2 do not, we can attribute this similarity to the estimation of subject-level effects by models in S_1.

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 an increase in estimate variability.

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

Model	Mod 0	Mod 1	Mod 2	Mod 3	Mod 4
Mod 0	0	-5.22	-5.73	83.4	191.0
Mod 1	5.22	0	-0.51	87.7	191.0
Mod 2	5.73	0.51	0	88.1	191.0
Mod 3	-83.4	-87.7	-88.1	0	179.0
Mod 4	-191.0	-191.0	-191.0	-179.0	0

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

Model	Mod 0	Mod 1	Mod 2	Mod 3	Mod 4
Mod 0	0	-36.0	-36.0	38.8	162.0
Mod 1	36.0	0	0	72.3	173.0
Mod 2	36.0	0	0	72.3	173.0
Mod 3	-38.8	-72.3	-72.3	0	146.0
Mod 4	-162.0	-173.0	-173.0	-146.0	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. Modle 2 to Model 3
    b. Model 0 to Model 4
2. The standard error decreases or remains constant on the following model transitions:
    a. Model 0 to Model 1
    b. Model 1 to Model 2

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 the incorperation of subject-specific correlation information into the variance component of the model.
c. The transitions in (2a) correspond to the incorperation of additive, subject-specific, predictor independent information into the model.
d. The transitions in (2b) correspond to the addition of information in the form of a subject-specific “Random Effect Intercept”

The preceeding relationships allow us to deduce the effects of the various types of information inclusion on our ability to make inferences on the realtionship between predictor and response. Beneficial information inclusions will result in reductions to standard error estimates (section 2 transitions, c & d relationships). Detrimental, or contradictory information will result in increase standard error estimates (section 1 transitions, a & b relationships).

The relationships outlined in (a)-(d) above are all based on the inclusion of various types of subject-specific information. These relationships can be classified as beneficial or detrimental to 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 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.

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.