Comparing Models of Subject-Clustered Single-Cell Data

Version 3.0

Lee Panter

Abstract

Single-Cell RNA sequencing data represents a revolutionary shift to the bioinformatic approaches being used to decode the human transcriptome. Such data are becoming more prevalent, and are being extended to multiple individuals, enabling analysis of subject-level relationships. However, it is not clear how to conduct this subject-level analysis. Current methods 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 single-cell RNA sequencing expression estimation in subject-correlated study design data. Each of the three methods: Linear Regression with Fixed Effects, Linear Mixed Effects Models with Random Effects, and Generalized Estimating Equations will have a detailed outline presented. We then compare the regression estimates and standard errors for each modeling method using real single-cell data from a Lupus Nephritis study of 27 subjects. We hoped that this paper presents insights into modeling single-cell expression data, and aids researchers with down-stream analyses.

Introduction

Traditional methods of sequencing the human transcriptome involve analyzing the combined genetic matrial of thousands or even millions of cells. These, so called "bulk", techniques are informative regarding population-average parameters, but often fail to capture the underlying variability in expression profiles within a sample population of genetic material [1].

Single-cell RNA sequencing (scRNA-seq) data sets are obtained by anylizing genetic material specific to individual cells. Hundreds or even thousands of cellular-specific genetic analyses performed on cells taken from within a single sample 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 genes that are differentially expressed across conditions [2]. Additionally, technological developments in whole-genome sequencing 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, test, and integrate 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. The modeling methods have been chosen so as to accomodate as much direct comparison of parameters across models as possible, while still altering the method by which subject-correlation is accounted for. We present model frameworks for each method, and compare the methods using subject-correlated, single-cell data from a study of 27 Lupus Nephritis cases.

Results Derived from Single-Cell Data

Before we begin a discussion of the data and methods used in this paper, we present several results established upon analysis of single-cell data over single-subject sources. It is hoped that the results presented will motivate similar analyses over multi-subject single-cell data sets.

Cellular sub-populations are often characterized by markers of differentiation that are limited to binary (present/absent) or discrete (eg. big, medium, small) values. The specific divisions resulting from these limited boundarys have lead to the common perception that the spectrum of human cell type is discretely delineated. This perception is being challanged in analyses of single-cell data that employ clustering methods such as Kohonen Self-Organized Maps [3] and visualization techniques such as ViSNE [4]. Stahlberg, Andersson, Aurelius, et.al [3] used Kohonen Self-Organizing Maps applied to single-cell data as a way of identifying rare cells within a homogenous population of neurological cells. Development of the ViSNE visualization method by [4] also demonstrated the utility of single-cell data in determining multi-dimensional boundary values for that distinguish healthy and cancerous bone marrow populations. The Kohonen Self Organizing Maps and ViSNE approaches to cellular sub-population classification using single-cell data are new and more robust compared to traditional methods of sub-population identifications. These methods search for markers of differentiation related to the observed trait of interest, and then assign a set of single-cell observations to that subpopulation that matches the differential marker criteria. Traditional methods of searching for markers of differentiation that first classify cellular subpopulations, then searching for the marker-subpopulation combination associated with the observed trait of interest.

Determining differential expression across condition is a method for (among others) researching a disease (condition), and identifying its genetic foundations. This genetic information can be useful in disease diagnostics, and treatment/prognosis once a diagnostic is made. Traditional

RNAseq methodologies that estimate population-level parameters will often fail to estimate the variability of expression profiles to fine enough resolutions to allow for identification of differential expression between condition groups. Conversely, Model-based Analysis of Single-cell Transcriptomics (MAST) [5] and Single-Cell Differential Expression (SCDE) [6] are mixture-model methods incorperating mean (positive) expression components, and zero-inflated (zero expression from technical or biological) expression sources. MAST and SCDE model single-cell specific information, and can therefore be used to find commonalities across observations specific to a condition. Both methods are suitable for determining differential expression across condition, and helping to demonstrate the versatility of single-celldata.

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" [7]. In this manuscript Arazi, Rao, Berthier, et al. looked to compare single-cell kidney tissue sample data from 45 Lupus Nephritis subjects vs. 25 population control samples [7]. The samples were collected from ten clinical sites across the United States, at which it was cryogenically frozen and shipped to a central processing facility. Samples were then thawed, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytomery [8]. The samples were then "dissociated", i.e. further prepared by disolving non-genetic, cellular and extracellular material. sc-RNA sequencing was performed using modified CEL-Seq2 method [9], followed by $\sim$ 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 [10] was used to examine intial data and perform quality control (QC) filtering of poor-quality observations. The Seurat package allows for easy classification of low-quality observations by setting threshold values for the following meta-data variables calculated automatically by the Seurat Package and independently verified:

1. $nFeature$ the number of unique genes detected in each cell
2. $PerctMT$ the percentage of reads that map to the mitochondrial genome

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

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

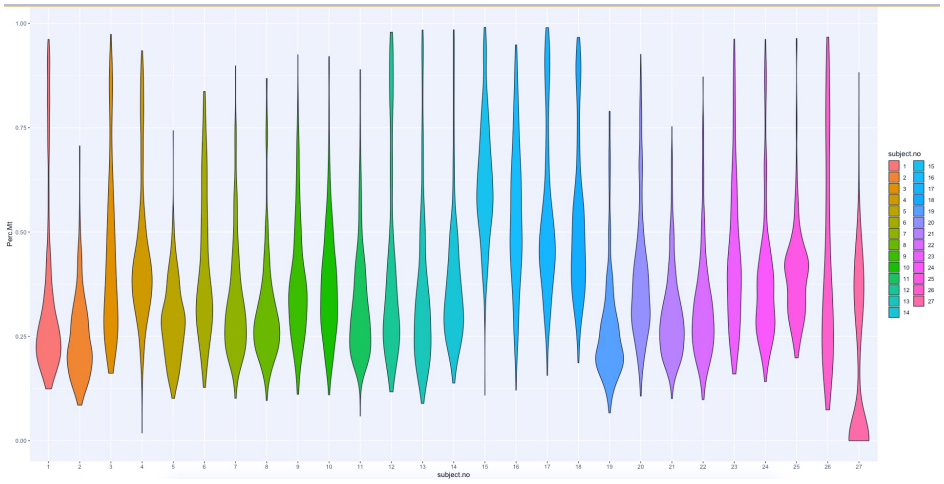

Figure 1:

The QC measures employed by Arazi, Rao, Berthier, et al. in [7] were:

1. $1,000 < nFeature < 5,000$

2. $PerctMT \leq 25\%$

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

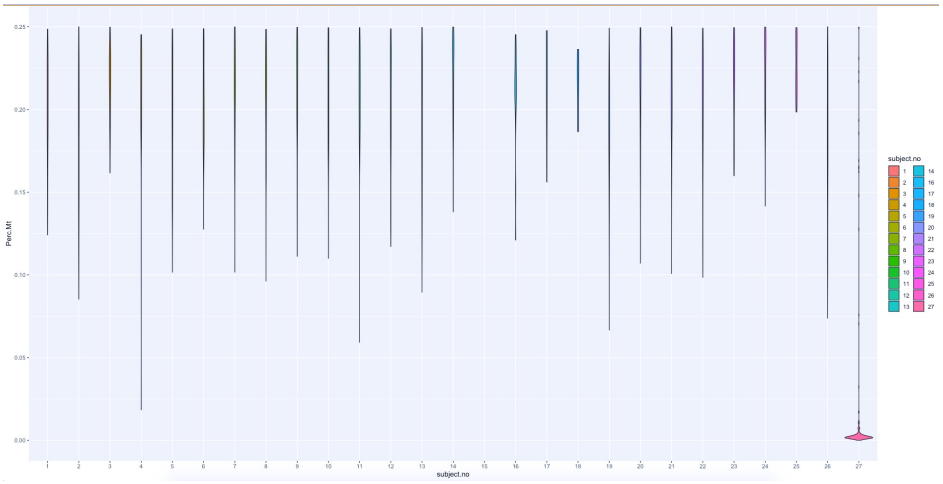

Figure 2:

a decision to increase the $PerctMT$ threshold to 60% was made to preserve the inherent distribution structure of data content where quality control restrictions could be met. 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 2) after filtering.

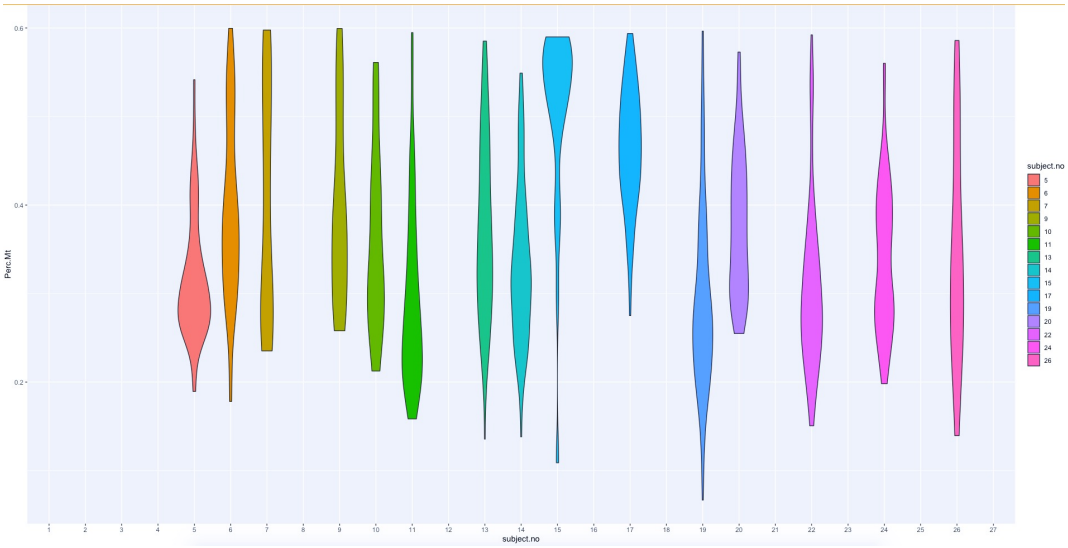

Figure 3:

The distribution of observations across subjects 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	0	0	0	0	58	86	32	0	31

Subject Number	10	11	12	13	14	15	16	17	18	19
Number of Observations	21	107	0	107	100	25	0	122	0	127

Subject Number	20	21	22	23	24	25	26	27
Number of Observations	75	0	87	0	79	0	53	0

Table 1

We note that the quality control process is an active population restriction, and the data being eliminated does not constitute "missing data" under the assumption that these values poorly represented the population of interest due to innacurate measurement. As a result, subjects which lack observations satisfying QC measures can be interpreted as non-informative as opposed to *missing* or *drop-out* events. This realization will allow us to reduce the data set distribution to informative subjects, for which the observational 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

Table 3 displays the minimum, median, mean, maximum, 1st and 3rd quartiles for the number of (non-zero) observations per subject:

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

Table 3

Variable Selection and Summaries

In order to simplify analysis and make more significant insights into model comparisons, we chose two pairs of variables from the 38,354 genetic markers in the Lupus Data to model in a predictor-response relationship. The variables we chose indicated higher values of correlation than arbitrary variable pairings, and are associated with conditions of interest (e.g. cancer treatment research in the case of MALAT1 [11], or observed limb malformations in the case of FBLN1 [12]). 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 Table (4) - (8). Each describes selected variable summary statistics (minimum, maximum, average, and median) for the subset samples specific to the subject identifiers used in Table (2).

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 4

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

Subject Number	Minimum	Maximum	Average	Median
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 5

CD134 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

Subject Number	Minimum	Maximum	Average	Median
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 6

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 7

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)

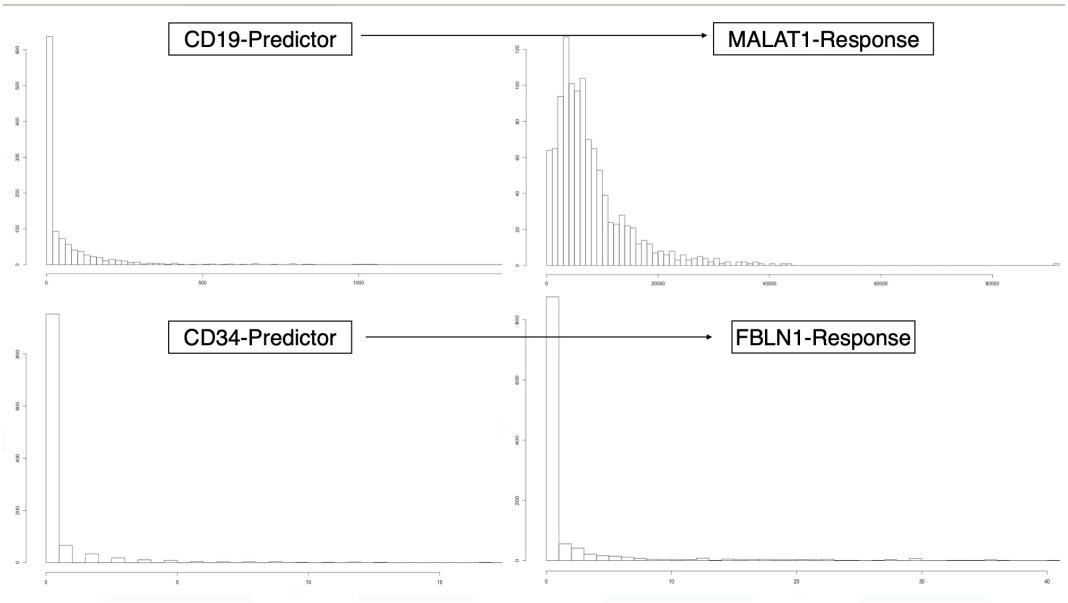

Figure 4:

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):

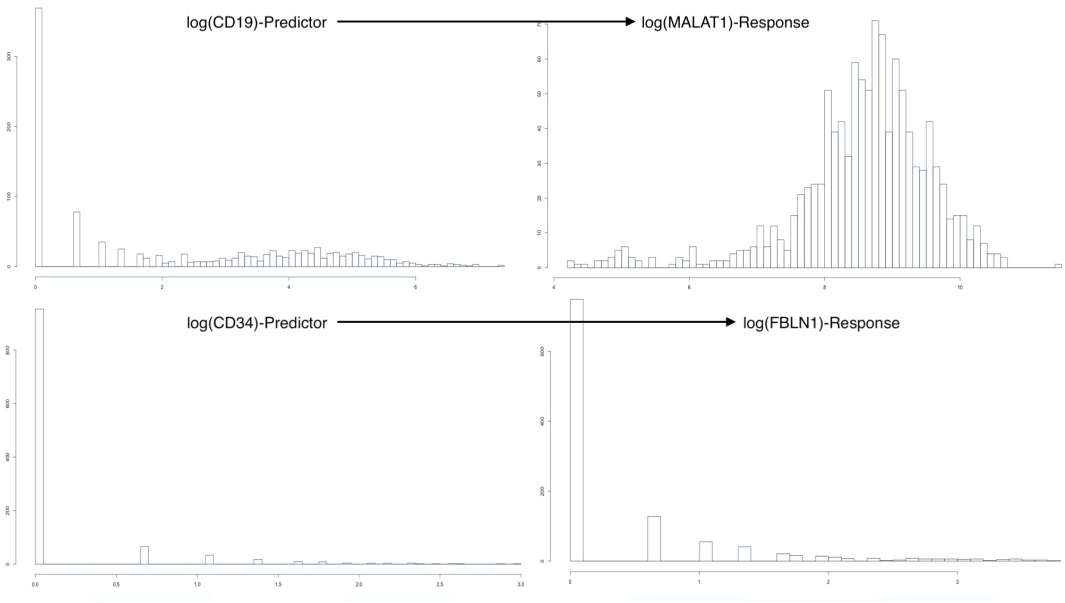

Figure 5:

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:

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

where

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

We aslo define the predictor attached to R_k as:

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

where

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

Let a single response be designated as: R_{hij}. The index $i \in \{5, 6, \ldots, 26\}$ 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 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 three 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 subject specific intercept and slope terms.

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.

Fixed-Effect Subject-Intercept (Model 1)

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

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

where we define the term:

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

Fixed-Effect Subject-Slope (Model 2)

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

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

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

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 observational independence. Correlation structures such as AR(1), independence, spatial power, or unstructured (lack of structure) 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 [13].

Linear Mixed Effects Model with Random Intercept (Model 3)

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 individual-specific observations, and attributing this source of variation to the nested sampling method. Such a model may be written as:

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

where

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

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

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 4)

Model 2 (Linear Regression with Fixed Effect Slope) implements a Fixed Effect slope in an attempt to reconcile the effects of observational correlation that was inadequately accounted for by the subject-specific Fixed Effect Intercept in Model 1. However, in light of the

information surrounding the development of Model 3, it is incumbent for us to develop an analogous correction for Model 2. Such a correction will allow us to account for observational correlation due to subeject-clustering 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 5)

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,

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

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 [13] 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 six modeling approaches:

(MALAT1 ~ CD19)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	4.918e-2	1.455e-2	3.381	7.47e-4
Model 1	4.833e-2	1.381e-2	3.500	4.84e-4
Model 2	5.143e-1	6.017e-2	8.546	< 2e-16
Model 3	4.920e-2	1.374e-2	3.579	3.6e-4
Model 4	5.938e-2	3.538e-2	1.678	1.19e-1
Model 5	4.92e-2	3.97e-1	1.53**	2.2e-1**

Table 8

(FBLN1 ~ CD34)

Model Number	Estimate	Std. Error	t-Stat	p-value
Model 0	7.884e-1	4.92e-2	1.6e+1	< 2e-16
Model 1	1.306e-1	3.42e-2	3.82	1.4e-4
Model 2	8.38e-2	5.89e-2	1.42	1.5492e-1
Model 3	1.35e-1	3.42e-2	3.95	8.4e-5
Model 4	1.705e-1	7.29e-2	2.34	6.7e-2
Model 5	7.88e-1	2.2e-1	1.281e+1**	3.4e-4**

Table 9

Note: ** These are Wald test of a single parameter (not t-tests)

The main-effect slope is of primary interest because of its interpretation as "the average relationship between predictor and response". In this context, we are able to interpret the magnitude and direction of main-effect slope parameter estimates using percentage change in the predictor as associated with multiplicative effects in the outcome. This interpretation can give intuitive meaning to the predictor-response relationship, and can be compared across models as different levels of subject-correlation are taken into account.

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 evidence strength the method is using to support its result. In other words, an increase in standard error between two models that are estimating the same parameter indicates an increase in estimate variability (a loss of precision).

It is worthwhile to note the consistency of estimates that were obtained. While not necessarily unexpected, the direction and magnitude of estimates and standard errors are largely comparable within and between variable pairings. An exception to this behavior are those estimates being generated by Model 2. In each of the variable parings, the estimate created by Model 2 is an order of magnitude off.

However, if we compare the remaing 5 models (exluding Model 2), we see that Models 0 and Model 5 have produced estimates that are more consistent with each other than the other methods-in both variable parings. This is an expected result as both Model 0 (Simple Linear Regression) and Model 5 (Generalized Estimating Equations) produce population-averaged estimates.

Changes in standard errors also display consistency properties. In each variable pairing:

1. The standard error increases on the following model transitions:
    a. Modle 3 to Model 4
    b. Model 0 to Model 5
2. The standard error decreases or remains constant on the following model transitions:

    a. Model 0 to Model 1
    b. Model 1 to Model 3

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.

Code and Data

All code for the above analysis was written and evaluated in RStudio Version 1.2.1335, and is available for download at the following GitHub repository:

https://github.com/leepanter/MSproject_RBC.git

Additionally, a link to all necessarry and referrence data files (including original data) are contained in the following Google Drive:

https://drive.google.com/open?id=1gjHaMJG0Y_kPYWj5bIE4gRJU5z9R2Wqb

References

1. Macaulay IC, Voet T (2014) Single cell genomics: Advances and future perspectives. *PLoS genetics* 10: e1004126.

2. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* 17: 63.

3. Ståhlberg A, Andersson D, Aurelius J, et al. (2010) Defining cell populations with single-cell gene expression profiling: Correlations and identification of astrocyte subpopulations. *Nucleic acids research* 39: e24–e24.

4. Amir E-aD, Davis KL, Tadmor MD, et al. (2013) ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology* 31: 545.

5. Finak G, McDavid A, Chattopadhyay P, et al. (2013) Mixture models for single-cell assays with applications to vaccine studies. *Biostatistics* 15: 87–101.

6. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nature methods* 11: 740.

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

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

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

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

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

12. 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.

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