Comparing Models of Subject-Clustered Single-Cell Data

Lee Panter

Audrey Hendricks, PhD-Committee Chair and Advisor[[1]](#footnote-1)

Stephanie Santorico, PhD-Committee Member[[2]](#footnote-2)

Rhonda Bacher, PhD-Committee Member[[3]](#footnote-3)

# Abstract

Single-Cell RNA sequencing data represents a revolutionary shift to approaches 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 only or both 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.

# 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].

Conversely, single-cell RNA sequencing (scRNA-seq) obtains 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 only as random intercepts (LMM-RI) or as both random intercepts and random slopes (LMM-RS), and generalized estimating equations (XXX). I first present the overall framework for each method. Then I compare the results for each model using single-cell data from a study of 27 Lupus Nephritis cases.

# Description of Data Set

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](https://www.immport.org/shared/study/SDY997).

MORE INFORMATION ABOUT THE ACTUAL DATA SET YOU HAVE IS NEEDED INCLUDING THE FOLLOWING:

* SAMPLE SIZE
* SAMPLE CHARACTERISTICS (I.E. CASES OR CONTROLS???)
* For example, the publicly available data that I use here consists of XXX cases???.

## 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, and ). 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. 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. 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 for each subject is displayed in (**Figure 1**) below:

A picture containing fence

Description automatically generated

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

All observations for which the calculated values of and satisfied the inequalities in (1) and (2) above were kept, and the others were considered “low-quality” and removed. The resulting distribution of the variable is displayed in (**Figure 2**):

A screenshot of a social media post

Description automatically generated

As 84% of cells as removed with the filters chosen by Arazi et al, I choose a more lenient threshold, removing observations with , 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 is displayed in (**Figure 3**) after filtering.

A picture containing clock

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## 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 I 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 zero-inflation components that are associated with protocol variations, and measurement error. Together, these factors contribute to right-skewed variable distributions (**Figure 4**)

A screenshot of a social media post

Description automatically generated

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

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):

The resulting distributions are shown in (**Figure 5**)

A screenshot of a social media post

Description automatically generated

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 model using the following notation for a subject level response and predictor pair:

where is an indicator for N subjects, and is an index for each cell for subject .

## Linear Model (LM)

The linear model canbe written as:

This model does not account for correlation structure in the data, and instead assumes the observations are independent. Linear model parameter estimates are for the population averages. The error term, , is assumed to be a normally distributed random variable with mean zero and variance .

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

Adding a subject-specific term to the LM model allows for the accounting of subject level effects by uniformly shifting the mean of the fitted values specific to a subject. This model is written as:

where

This model adds estimated parameters which represent the average deviation for each subject from the global estimated mean Linear Model (LM).

## Linear Mixed Effects Models

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 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. 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. The LMM-RI model is written as:

where

it is assumed that and 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. 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:

where

## Generalized Estimating Equations (GEE)

Generalized Estimating Equations (GEE) estimates 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 , where

And

A screenshot of a cell phone

Description automatically generated

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

A picture containing object, clock

Description automatically generated

The first derivative matrix is given by:

is the working covariance matrix. This matrix can be estimated from hyper parameters created in the process of numerically solving for the GEE estimates that solve . Please see Fitzmaurice, Laird, and Ware [9] for more information.

The GEE algorithm uses the following general steps to converge at an estimate:

1. Maximum likelihood estimation is used are used to obtain intial estimates for
2. Estimates for (from 1) used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix () are used to obtain new estimates for 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 an estimate of the average population.

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

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

The three-part specification of the GEE framework includes:

1. The link function and linear predictor
2. Variance function
3. A working covariance matrix
   1. Here, the identity link function is used:

in conjunction with the linear predictor:

which implies the modeling equation:

* 1. An identity variance function is used
  2. 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:

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

An interpretation of this parameter is: **across all subjects, a one-unit increase in the predictor () is associated with a unit change in the expected outcome ()**

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 represents an estimate obtained for the fixed effect slope as obtained by one of the previously describe modeling methods. An interpretation of this parameter is: **for a given subject (controlling on the subject of origin) a one-unit increase in the predictor () is associated with a unit change in the expected outcome ().**

# Results

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Displayed in (**Table 5)** – (**Table 10**) below are percent change in: parameter estimate, standard error, and test statistic for the variable pairing in **Tables (5)-(7)** and the variable paring in **Tables (8)-(10)** . Where the percent change in estimate is defined as:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## A screenshot of a cell phone Description automatically generated

## A screenshot of a cell phone Description automatically generated

A screenshot of a cell phone

Description automatically generated

## A screenshot of a cell phone Description automatically generated

## 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 similar down to . The LM-FE and LMM-RI method estimates are also similar since estimates for each estimate (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 smallest 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 larger p-values up to three orders of magnitude of percent change.

## Nested Model Comparisons

A screenshot of a cell phone

Description automatically generated

(**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.

A screenshot of a cell phone

Description automatically generated

(**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.

A screenshot of a cell phone

Description automatically generated

(**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 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 variable pairing.

# Discussion

Here, I compared five modeling strategies for detecting subject level associations in single-cell RNA sequencing data gathered over 27 subjects from a Lupus Nephritis study: 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 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 normality of the errors, 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.

The analyses performed here are 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

A close up of a device

Description automatically generated

## Table A2

A screenshot of a cell phone

Description automatically generated

## Table A3

A close up of a device

Description automatically generated

## Table A4

A close up of a device

Description automatically generated

# 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.

1. The University of Colorado-Denver [↑](#footnote-ref-1)
2. The University of Colorado-Denver [↑](#footnote-ref-2)
3. The University of Florida [↑](#footnote-ref-3)