Results

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

Coefficient Estimates

(MALAT1 ~ CD19)

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

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

(FBLN1 ~ CD34)

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

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

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

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

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

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

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

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

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

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

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

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

Standard Error Estimates

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

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

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

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

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

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

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

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

a. The modeling transitions in (1a) correspond with the addition of information to the model in the form of a subject-specific "Random Effect Slope".
b. The transitions in (1b) correspond to either:
    i. addition of the parameter in 1a
    ii. loss of subject-specific information that was originally incorporated into the variance-component of the model. I.e., loss of subject-specific variability information.
c. The transitions in (2a and 2b) are the inverse representation of the relationships outlined in (a) and (b) above.
d. The transition in (2c) corresponds to the incorporation of additive, subject-specific, predictor independent (mean-effect) information into the model.
e. The transition in (2d) corresponds to the addition of the assumption of independence between subjects, along with a purely parametric fitting method (for LM-FE).

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

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

Test Statistics

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

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

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

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

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

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

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

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

The results outlined in the section above are all based on the inclusion of various types of

subject-specific information. These relationships can be classified according to how they affect our ability to perform inference on the relationship between a predictor and a response using subject-correlated scRNA-seq data. To this effect, we can now evaluate our variable-pairing relationship(s) to determine if there is a significant effect from the nested sampling methods used to create the scRNA-seq data, and if there is an effect, how can this effect best be accounted for.

Discussion

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

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

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

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

Although the Seurat Guided Clustering Tutorial [6] provides a framework for quality control with integrated exploratory analysis, the observed protocol dependencies of scRNA-seq data must still be considered before any analysis can be conducted. While methods of combining existing scRNA-seq data have been used to successfully integrate multiple-subjects' single-cell observations [10], no batch-effect corrections or expression normalization has been performed to account for sources of possible confounded or misrepresented subject-level correlation effects.

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