# Discussion

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

I find that population average models (i.e. LM and GEE) and subject specific intercept models (i.e. LMM-RI and LMM-RS) tend to produce similar results within the same model class (population average or subject specific intercept) but different results between model classes. The highest standard errors are indicated in the LMM-RS model, and the lowest standard errors in the LMM-RI model. LM-FE standard error is also found to be smaller than both LM and GEE standard error values. Nested model comparisons indicate that inclusion of subject specific terms is advisable at all levels (fixed and random, intercept and slope) with exception of the random slope in the $FBLN1 \sim CD34$ variable paring.

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

This analysis is subject to several drawbacks and limitations. All the results are based on evidence obtained from just two single-cell RNA sequencing variable pairings. In the future, comparing the consistency of these models over all model pairs is needed. Additionally, single-

cell RNA sequencing data is heavily influenced by protocol dependencies and measurement inconsistencies. Quality control must be carefully considered and conducted prior to any analysis.

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