Comparing Models of Subject-Clustered Single-Cell Data

Version 6.0-Discussion

Lee Panter

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 [1] 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 [2], 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.

References

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

2. Stuart T, Butler A, Hoffman P, et al. (2019) Comprehensive integration of single-cell data. *Cell* 177: 1888–1902.