Comparing Models of Subject-Clustered Single-Cell Data

Abstract

Single-Cell RNA sequencing data represents a revolutionary shift to approaches being 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.