Final Project

Phase 1 – Initial Project Description

Lee Panter

Data Description

The data I am proposing to work with is single-cell resolution measurements of RNA sequencing and Flow Cytometry variables taken across 27 subjects. Each of the 27 subjects has an average of 354 "repeated measures", on variables that are: continuous (in the case of Flow Cytometry measurements), count-valued (in the case of RNA sequencing measurements), and class/factor-valued (in the case of meta-data associate with each observation). There are 19 Flow Cytometry variables measured, and 38354 unique genetic markers identified within the SmartSeq RNA sequencing process (Arazi et al. 2018). There are also 15 meta-data variables. Accounting for all of these observations, and allotting for redundancies across merged data sets, there will be $\sim 3.66 * 10^8$ individual measurements for which to account.

Key Research Question(s) of Interest

What modeling methodologies would allow for the most stable and accurate subject-level parameter inference(s) on quality-controlled and cellularly homogeneous data?

Specifically, we look to compare subject-level parameter estimates made from simple models built on single or (at most) two covariate subsets, with fixed or random subject effects, modeling various output variable types from a quality-controlled and homogenized data subset.

Explain How the Data are Correlated

The experimental techniques employed in (Arazi et al. 2018) would indicate that:

- subject-level measurements are made independently
- cellular-level measurements within each subject are made independently

It would be my preference to proceed with this assumption, or to perform tests for independence on randomized subsets. In the event that the independence hypothesis is rejected (at either level), a different structure could be assumed; however, given the complexity of the data measurements, the covariance structure might need to be simplified for computational reasons.

What Makes the Dataset Interesting?

This data set is interesting (to me) because of the different measurement outcome types (categorical, count, and continuous), and because the research question of interest, if well answered in a concrete fashion, could be helpful scientists in an active area of research.

What Makes the Dataset Messy or Unique?

I have been working on Quality Control for this data for almost two months. These two months have given me extremely valuable insight into how the experimental design directly impacts the RNAseq data, and why this could be VERY problematic.

The data suffers very badly from "$p >> n$" syndrome, and finding important variables in the noise is challenging.

The observations are unbalanced between subjects, and the RNAseq data seems to be imbalanced across certain subjects (which will make subject-level inferences significant and interesting).

References

Arazi, Arnon, Deepak A Rao, Celine C Berthier, Anne Davidson, Yanyan Liu, Paul J Hoover, Adam Chicoine, et al. 2018. "The Immune Cell Landscape in Kidneys of Lupus Nephritis Patients." *bioRxiv*. Cold Spring Harbor Laboratory, 363051.