Perturb Seq

Caleb Lareau

March 19, 2017

Introduction

In three recent (December 2016) papers published in the journal *Cell*, researchers at the Broad Institute, UCSF, and the Weizmann Institute described a new technology called Perturb-Seq, which integrates CRISPR-Cas9 genome editing and droplet-based single-cell RNA-Seq. **Figure 1** provides a graphical overview of this technology.

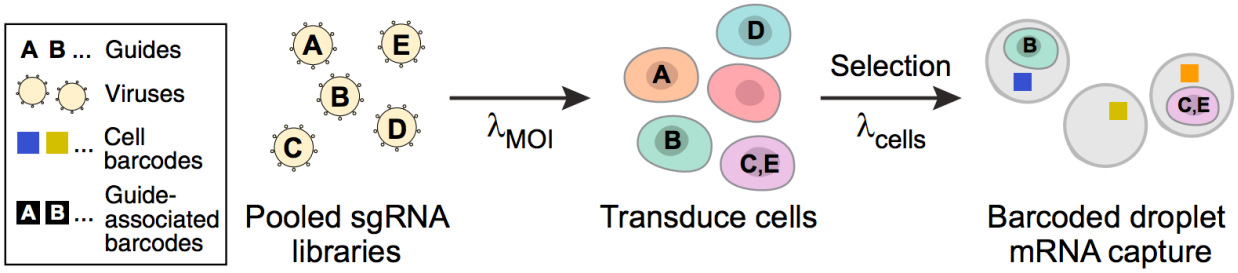

Figure 1: **A graphical overview of the experimental design of Perturb-Seq.** In brief, viruses containing different genome-editing vectors (labeled **A,B,C,D,E**) infect cells and then modify the host cells' DNA using CRISPR/Cas-9. These modified cells are then experimentally examined using a high-throughput microfluidics droplet-based capture. The result of this experiment are data with a complex correlation structure that I propose to be examined further.

The basic idea is to knock-out ("Perturb") certain key regulators (equivalently, genes or transcription factors) in a cell and **then examine how the cell responds to these changes.** The cellular response is measured through RNA-**Sequencing**, hence the name of the technology. Though the biological underpinnings of Perturb-Seq in itself is not new, the **sample size** ($> 200,000$) and **resolution** (single cell) distinguish this method from all other existing technologies. In essence, Perturb-Seq is doing thousands of carefully controlled experiments in parallel, enabling an unprecedented throughput of experimental data. Early review articles have suggested that Perturb-Seq could be **the key technology for dissecting gene regulation in the context of disease**. So far, only a handful of primarily biologically-motivated folk have begun to think about modeling these very complicated data. **The goal of this project, thus, is to 1) convey the importance and theory of this experimental data source, 2) discuss existing modes of estimation/inference, and 3) situate the analyses problems in the context of discussion of BST 245**.

State of the stats

Atray Dixit maintains a Github repository (https://github.com/asncd/MIMOSCA), which contains a python script with about 2,000 lines of code that defines the backbone of the statistical methodology for Perturb-Seq. In brief, this code base fits an elastic net linear model to a design matrix and gene expression readout as

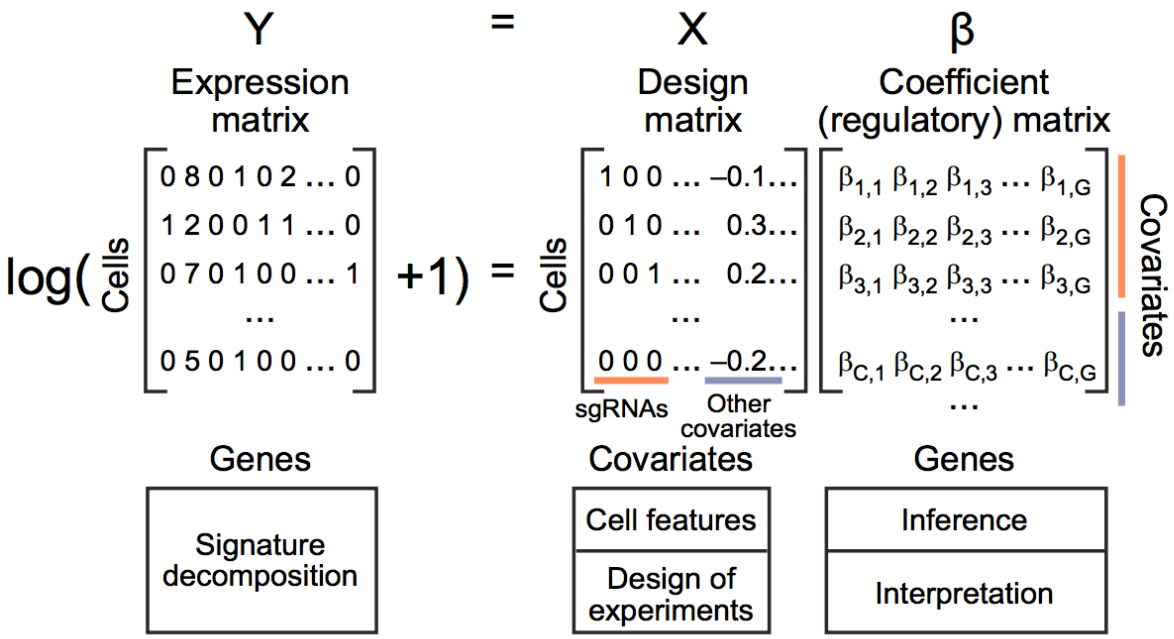

Figure 2: **A graphical overview of the generalized linear model fit in Perturb-Seq analyses.**

depicted in **Figure 2**. Of note, the dimensions for this analysis will be on the order of 50,000 cells, 30,000 genes, and 50 covariates. **There is no existing infrastructure to examine this data in `R`, which will be a component of the "resource" uility in the proposed project**. The aim/expectation is that "translating" the presentation of the data from the authors into a format more typically consumed by the Biostastics department will expedite statistical innovation that is direly needed for this data type.

Links to the BST245 Course

- Repeated measurements (the same cell/sgRNA combination is detected $\sim$ 500 times)
- Correlated independent variables (effects of sgRNAs may be correlated if they underlie the same regulatory process)
- Correlated dependent variables (gene RNA values are inherently highly correlated)

Presentation Outline

- 0-7 minutes: Discussion of CRISPR/Cas9, scRNA-Seq, Drop-Seq, and gene regulation
- 7-12 minutes: Discussion of Perturb-Seq
- 12-15 minutes: Key takeaways of the Perturb-Seq paper as presented
- 15-20 minutes: Translating between Perturb-Seq and BST245 Notations
- 20-25 minutes: High-level exploratory data analysis
- 25-32 minutes: Discussion of model fits previously implemented
- 32-38 minutes: Analysis of model fits, coefficients, etc.

- 38-40 minutes: Immediate recommendations for improvement of statistical models
- 40-50 minutes: Discussion

References

Adamson, Britt, et al. "A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response." Cell 167.7 (2016): 1867-1882.

Dixit, Atray, et al. "Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens." Cell 167.7 (2016): 1853-1866.

Jaitin, Diego Adhemar, et al. "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq." Cell 167.7 (2016): 1883-1896.

Wagner, Daniel E., and Allon M. Klein. "Genetic screening enters the single-cell era." Nature Methods 14.3 (2017): 237-238.