Allele-specific expression from single-cell RNA-Seq data

Kwangbom Choi, Narayanan Raghupathy, Steven C. Munger, Gary A. Churchill
The Jackson Laboratory, Bar Harbor, Maine 04609, U.S.A.

Background

In diploid cells, two allelic copies of a gene can differ in the timing and the level of their expression as determined by genetic, environmental, and stochastic factors. With high-throughput sequencing (HTS) technologies, we can now resolve how alleles are preferentially expressed in tissue samples and in individual cells. This information not only reveals which alleles are preferentially expressed, but enables us to decode how gene expression is regulated. This insight is fundamental for understanding the genetic architecture underlying normal phenotypic diversity and disease.

Although the application of single-cell RNA-Seq methods (scRNA-Seq) adds valuable information about the dynamics of allelic expression that is lost in whole tissue samples, it poses multiple technological challenges. High level of sampling noise is common, often accompanied by relatively low depth of coverage (below 10 million reads per cell). Due to inefficient, non-random reverse transcription process, alleles may drop out from measurements. Accurate quantification of allele-specific expression (ASE) from scRNA-Seq data requires allelic variation to distinguish reads from different alleles, but many reads will not overlap polymorphic sites and their origins are ambiguous.

We propose an empirical Bayes model in which we disambiguate the origins of multiply-aligning reads (or multireads), quantify ASE in individual cells using all the aligned reads, refine ASE in each cell referring to other cells in similar expression state, and classify genes by summarizing how cells behave across the population on their transcription events.

Results

Relying solely on uniquely-aligning reads was not a viable option for quantifying ASE from scRNA-Seq data. In many cases, over 80% of reads had to be filtered out just because they aligned to multiple locations of diploid transcriptome. Discarding multireads increases the variability of ASE across genes and results in false discoveries, for example, hundreds of false monoallelic expression patterns. Our model also found strong evidence on coordinated expression of alleles in many genes: gene-specific odds of joint expression of two alleles suggests positive correlation. We were able to classify genes into seven categories with respect to allelic expression state of cell population: maternal monoallelic, paternal monoallelic, biallelic expression, and four combinatorial mixture of those three base classes, including mutually exclusive allelic expression. The classifier helped us, for example, to identify genes dynamically change their expression state along the progress of tissue development.

Conclusions

Our model overcome data sparsity in each individual cell by combining information from other cells of a kind, and control overdispersion by allowing cellular heterogeneity on allele proportion via a hierarchical model. We also provide a heterogeneous model to describe sub-populations of cells resulting from random mono-allelic expression in scRNA-Seq data, which thus enable us to derive shrinkage estimation on allele specificity out of cells in diverse expression states.