Final Project

Mallory Lai

November 17, 2016

Although several statistical methods have been used to examine RNA-Seq count data, no standard method exists. baySeq is an R package that analyzes RNA-Seq data in a Bayesian context. By default, the method assumes a Negative Binomial (NB) distribution as its prior and calculates a posterior probability that each gene is differentially expressed. Although it's known to have a low False Discovery Rate (FDR), it's computationally intensive and relatively conservative.

The goal here is to investigate just how conservative such analyses are since error and assumptions made at this stage can greatly affect downstream analyses. The project's aim is to quantify the proportion of accurately predicted differentially expressed genes using baySeq using simulated RNA-Seq count data. I will use the default Negative Binomial distribution as well as the user-specified Zero-Inflated Negative Binomial (ZINB).

Methods

Simulated RNA-Seq data was produced using the compcodeR package and differential gene expression was analyzed using the baySeq package. Two hundred simulations were performed in total. Half of them were analyzed with a Negative Binomial prior in baySeq and the other half with a Zero-Inflated Negative Binomial. Total computation time was ~48 h with parallel processing using the parallel package in R.

The compcodeR package was used to produce simulated RNA-Seq count data for two groups of five samples each. Counts are simulated from a Negative Binomial distribution with equal dispersion between the two groups. Simulated data was automatically filtered to exclude genes with values of zero among all five samples. No outliers were introduced.

For each simulation, a random number of genes between 8,000 and 12,500 (the maximum allowed) was tested. A random percentage of those genes, between 5-30%, were selected to be differentially expressed. The fraction of differentially expressed genes to be upregulated was also randomized to be a proportion between .45 and .65.

Differential gene expression was then analyzed using the baySeq package. The default Negative Binomial prior was used for half of the simulations and a Zero-Inflated Negative Binomial was used for the rest. Differentially expressed genes were classified as those with a false discovery rate less than 0.05.

The proportion of accurately classified genes for each simulation were then .... PDF

A filtered PDF was created for the NB and ZINB groups. The filter width was varied to minimize noise without spreading the PDF out too much.

Results

Discussion

Further research should center on the accurate classification of differentially expression genes under a wider range of simulated conditions. Simulations with more than two groups and varying replicate numbers should be considered. The number of genes, the proportion of differentially expressed genes, and the fraction of upregulated genes should include a wider range of variables. Simulations which involve no differentially expressed genes and only upregulated or downregulated genes should be considered as well. Without such conditions, these results cannot be considered robust and generalizable to the conditions not tested.