Evaluating baySeq's differential gene expression analysis of RNA-Seq data with compcodeR

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 with simulated RNA-Seq count data. I will use the default Negative Binomial distribution as a prior 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. One thousand 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 ~80 h with parallel processing using the parallel package in R. Approximately 72 hours of computational time was performed on a Windows 10 desktop and ~8 hours was performed on the cluster computer Mt. Moran in Laramie, WY.

The compcodeR package was used to produce simulated RNA-Seq count data for two groups of five samples each. Counts were 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. Accurately classified genes were defined as those which correctly identified upregulation or downregulation with a posterior likelihood greater than .95.

The proportion of accurately classified genes for each simulation was then calculated by dividing the number of accurately classified genes by the number of differentially expressed genes defined by compcodeR. A filtered probability density function (PDF) of the proportion of accurately classified genes was created for the NB and ZINB groups. The filter width was chosen to minimize noise without spreading the PDF out too much.

A Beta PDF was fit to the data. The Beta distribution was used since the proportion of correctly classified genes is bounded between 0 and 1. The parameters α and β were calculated from the mean and variance of the data:

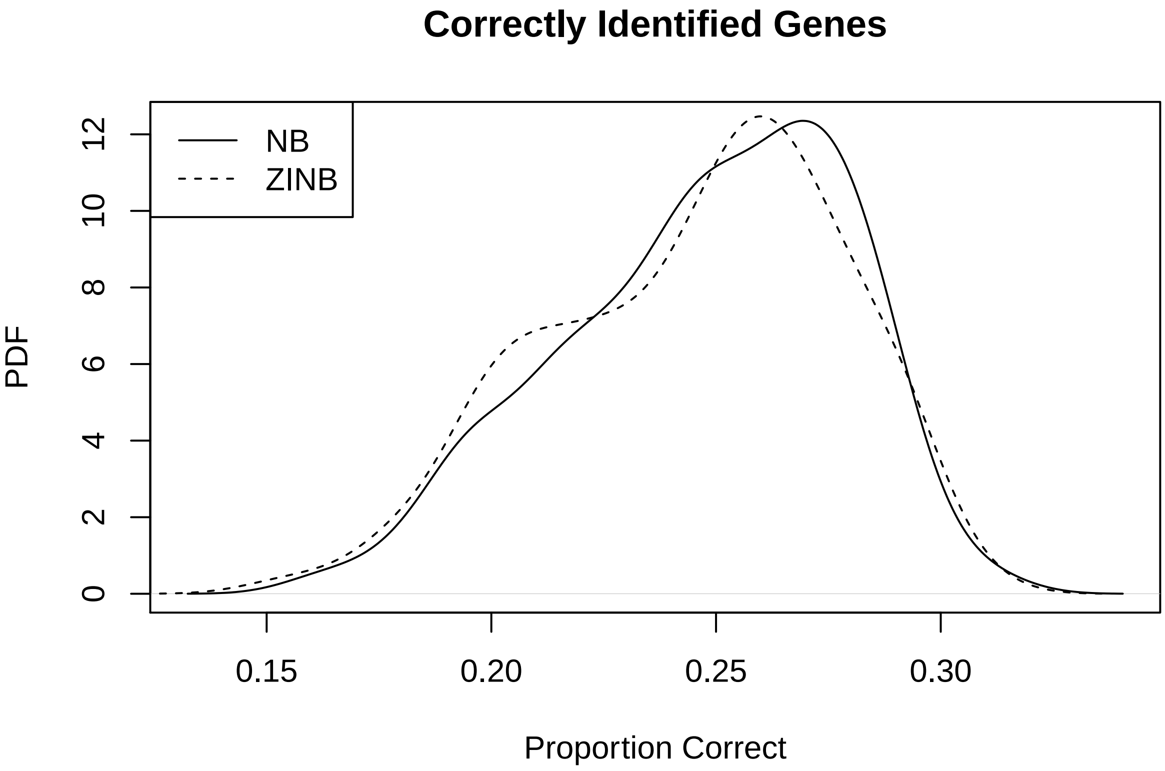
$$\alpha = \langle X \rangle \left[\frac{\langle X \rangle \left(1 - \langle X \rangle\right)}{\left\langle \tilde{X}^2 \right\rangle} - 1 \right]$$

$$\beta = (1 - \langle X \rangle) \left[\frac{\langle X \rangle (1 - \langle X \rangle)}{\left\langle \tilde{X}^2 \right\rangle} - 1 \right]$$

Results

No significant differences between the NB and ZINB priors were found. The mean for the NB prior was ________ and ________ for the ZINB prior. The variance was also comparable–_____ for the NB and ________ for the ZINB. Because no significant differences were found between the NB and ZINB priors, the data was combined to form one PDF.

	Mean	Variance
NB	0.2450512	0.0010727
ZINB	0.2476911	0.0009806



Discussion

Although the FDR for baySeq is indeed quite low, the proportion of accurately identified genes is low with a large computational cost. Other packages, such as edgeR or DEseq, although not Bayesian in nature, outperform baySeq in terms of both computational cost and number of genes detected.

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.