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 differential gene expression through 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, it's computationally intensive and relatively conservative in terms of identifying differentially expressed genes.

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 and its true positive rate using baySeq with simulated RNA-Seq count data through compcodeR. The default Negative Binomial distribution will be used as a prior as well as the user-specified Zero-Inflated Negative Binomial (ZINB) for comparison.

Methods

Simulated RNA-Seq data was produced using the compcodeR package and differential gene expression was analyzed using the baySeq package. Three thousand simulations were performed in total. Half of them were analyzed with baySeq using an NB prior and the other half with a ZINB prior. Total computation time was ~90 h with parallel processing using the parallel package in R. Approximately 72 hours of computational time was performed on a Windows 10 desktop and ~18 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. Differentially expressed genes (DE genes) from baySeq were classified as genes with a posterior likelihood greater than .95, whereas true DE genes were defined as the genes simulated to be upregulated or downregulated by compcodeR. The detection of a true DE gene by baySeq required the correct identification of upregulation or downregulation. A true positive rate was defined as the proportion of true DE genes out of the the total DE genes detected by baySeq. The detection rate is the number of true DE genes detected by baySeq divided by the total number of true DE genes determined by compcodeR, indicating the proportion of true DE genes baySeq was able to detect.

A filtered probability density function (PDF) of the detection rate and true positive rate 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 then fit to both PDFs. The Beta distribution was used since the proportion of correctly classified genes is bounded between 0 and 1. Parameters α and β were calculated from the mean and variance of the data with the following equations:

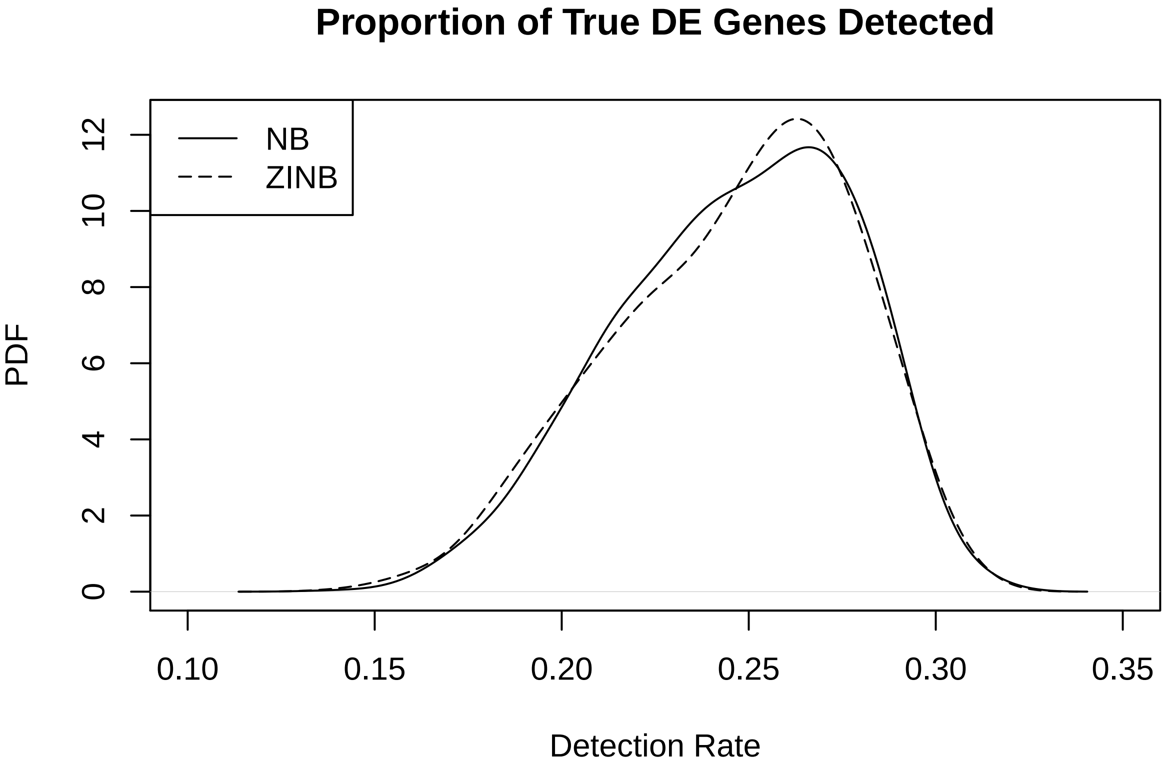
$$\alpha = \langle X \rangle \left[\frac{\langle X \rangle \left(1 - \langle X \rangle\right)}{\left\langle \tilde{X}^2 \right\rangle} - 1 \right]$$

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

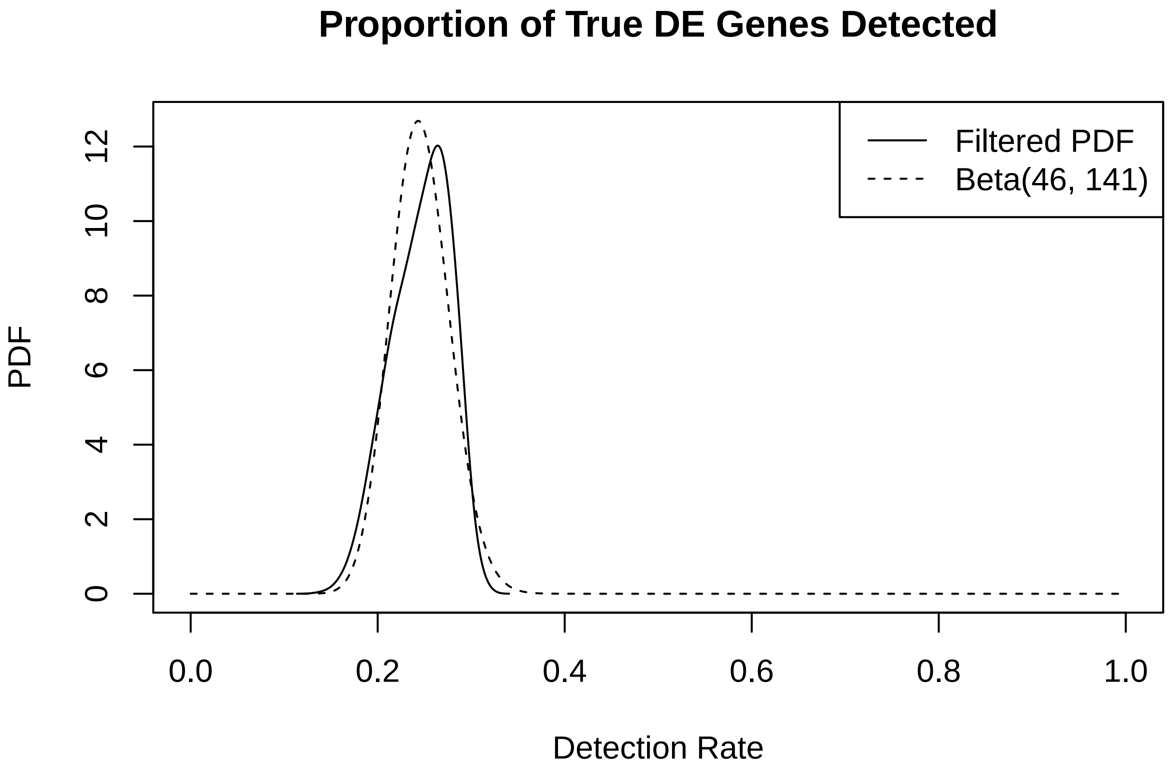
Results

Various filtered PDFs were created for the detection rate of both NB and ZINB groups. A filter width of 0.035 was determined to provide the best shape for the PDF–minimizing noise without spreading out the PDF too much. No significant differences between the NB and ZINB priors were found. Both the mean and variance was similar amongst both groups.

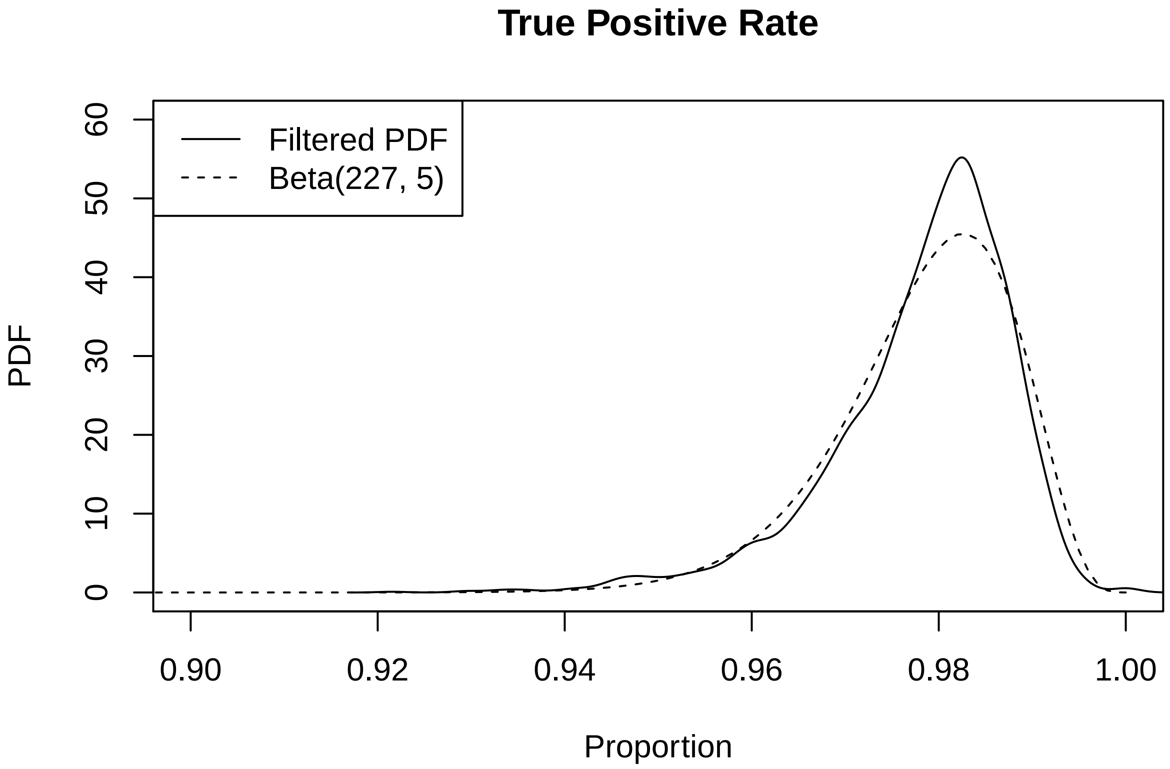
	Mean	Variance
NB	0.2459823	0.0010124
ZINB	0.2463144	0.0009650



Because no significant differences were found between the NB and ZINB priors, the data was combined to form one PDF. The parameters α and β were calculated from the combined output to yield a Beta PDF with $\alpha \approx 46$ and $\beta \approx 141$. The filtered PDF was plotted against the Beta(46, 141) PDF for comparison.



A filtered PDF was also created for the true positive rate for the combined NB and ZINB groups. Again, a filter width of 0.035 was found to be optimal. The average true positive rate was found to be ~98%. Here, the parameters α and β were calculated to yield a Beta PDF with $\alpha \approx 227$ and $\beta \approx 5$. The Beta(227, 5) and filtered PDF were plotted against each other for comparison.



Discussion

Although not perfect, the Beta PDFs describe the detection rate and true positive rate for baySeq reasonably well. The Beta(46, 141) slightly overpredicts detection rate near the filtered PDF's peak. The Beta(227,5) does the opposite–underpredicting the true positive rate near its peak. However, it is possible that with more simulations the filtered PDF may come closer to resembling the Beta PDFs.

Further research could 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 also 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.

Although the true positive rate of ~98% for baySeq is indeed quite high, the detection rate of ~25% 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 detection rate. It is possible that lowering the threshold for significance of the posterior likelihood could bring detection rates up while keeping the true positive rate within a reasonable limit. However, until detection rates can be improved, it may be better to use a less computationally intensive package for differential gene expression analysis.