**EBSeq: An empirical Bayes hierarchical model for identifying differentially expressed genes in RNA-seq experiments**

Ning Leng1,4, John A. Dawson2, James A. Thomson3, Victor Ruotti4, Anna I. Rissman2, Bart M.G. Smits4, Jill D. Haag5, Michael Gould2, Ron M. Stewart2 and Christina Kendziorski1,4

1Department of Statistics, UW-Madison, WI; 2Morgue Institute for Research, Madison, WI; 3Department of Oncology, UW-Madison, WI; 4Department of Biostatistics and Medical Informatics, UW-Madison, WI 53706

**ABSTRACT**

Messenger RNA expression is important in normal development and differentiation, as well as in manifestation of disease. High-throughput DNA sequencing (RNA-seq) experiments allow for the identification of differentially expressed (DE) genes and their corresponding isoforms on a genome-wide scale. However, statistical methods are required to ensure that accurate identifications are made.

A number of methods have been developed for identifying DE genes in RNA-seq experiments, but they are deficient for identifying DE isoforms. Because uncertainty in estimated isoform expression varies directly with isoform complexity, applications of gene-centric isoforms and applications of gene-centric isoforms need to be considered in the two condition comparison, as we will see below.

**METHODS**

**Isomorph Simulation With Outliers**

We follow a simulation setup similar to that in the Robinson and Smyth (2007) (edgeR) paper with the assumption that isoform counts within condition are independent, in which N is sampled from the empirical ones at each condition and n is sampled from the empirical ones within each condition. This is performed using isoform simulation software.

DESeq, edger, and baySeq are applied to all of the isoforms at once, and then to each N group individually.

**Simulation Results**

**Gene Simulation With Outliers**

The first gene level simulation set-up is similar to the isoform simulation, but with T sampled from the empirical ones at gene level. The above shows that in the absence of outliers, DESeq and edger had the highest power for identifying DE genes. However, the FDR was increased almost three-fold. EBSeq was robust, showing the highest power among these methods with well controlled FDR.

To further investigate the relationship between outliers and DE identifications, we used the multiple condition EBSeq model on the combined data set with 24 samples to evaluate the posterior probability of each of 18 patterns of expression.**

**CONCLUSIONS**

The identification of isoforms DE across two or more biological conditions is a common and important problem in RNA-seq experiments that is often addressed by applying statistical methods for identifying DE genes directly to isoforms. As gene-based methods do not account for differential uncertainty inherent in isoform expression estimation, their application leads to underpowered inference for some classes of isoforms and inflated false discoveries for others. To address this, we have developed an empirical Bayesian hierarchical modeling approach, EBSeq, that enables accurate, efficient, and robust inference for isoform-seq experiments and is more robust to outliers on both gene-level and isoform-level inference. The manuscript could be found at UW-Madison BIMM technical report no. 226.

**Contact Information:**

Ning Leng

christina.kendziorski@biostat.wisc.edu

**Acknowledgement:**

This research was supported by National Institutes of Health (NIH) grants R01 GM085216 and R01 GM081823, and the Department of Defense (DoD) award W81XWH-12-1-0303.