Statistical Issues in the Analysis of Data from RNA-Seq Experiments

David M. Rocke*
Division of Biostatistics and Department of Biomedical Engineering
University of California, Davis, CA, USA – dmrocke@ucdavis.edu

Luyao Ruan
Division of Biostatistics and Department of Biomedical Engineering
University of California, Davis, CA, USA – lruan@ucdavis.edu

Blythe Durbin-Johnson
Division of Biostatistics
University of California, Davis, CA, USA – bpdurbin@ucdavis.edu

Sharon Aviran
Department of Biomedical Engineering
University of California, Davis, CA, USA – saviran@ucdavis.edu

RNA-Seq data are increasingly used for whole-genome differential mRNA expression analysis in lieu of gene expression arrays such as those from Affymetrix and Illumina. We review commonly used methods for this type of analysis, including DESeq, edgeR, and limma-voom, by placing them within a common framework that allows comparisons of components of the methods as well as of the overall results. We also review a number of recent studies comparing these methods in terms of false positives and sensitivity, and add additional results of our own. We show that many of the commonly used methods are not satisfactory, with most identifying large numbers of genes as differentially expressed even when there are none. This area is still early in its intellectual development and is changing rapidly, so there are substantial contributions that can be made.

Keywords: RNA-Seq; Gene Expression; Negative Binomial; DESeq; edgeR; limma-voom.