Controlling False Discovery Rates in RNA-Sequencing Data

Conrad J. Burden*
Australian National University, Canberra, Australia
conrad.burden@anu.edu.au

Sumaira Qureshi
Australian National University, Canberra, Australia
sumaira.qureshi@anu.edu.au

Susan R. Wilson
University of New South Wales, Sydney, Australia and Australian National University, Canberra, Australia sue.wilson@anu.edu.au

High throughput sequencing technologies are supplanting microarrays as the preferred technology for detecting and quantifying differential gene expression. The raw data produced by the a technique known as RNA-sequencing (RNA-seq), consists of integer counts of reverse transcribed cDNA fragment reads mapped onto each gene or transcript isoform in a reference genome or transcriptome. Many software packages exist for analysing RNA-seq datasets consisting of tables of mapped read counts from biological or technical replicate experiments under two or more conditions, the purpose being to detect which genes are differentially expressed between conditions. Two state-of-the-art packages, DESeq and edgeR, are based on a negative binomial model of read counts.

Our tests with simulated data constructed according to the statistical model assumed by these packages reveal that both packages generate a non-uniform p-value spectrum from null-hypothesis data. We demonstrate how specific knowledge of the non-uniformity can be exploited to develop a graphical technique based on the Storey-Tibshirani method for improving estimates of p-values and false discovery rates in databases where differential expression is present. We have developed an add-on package for DESeq and edgeR, called Polyfit, which implements this method, and evaluate its performance against DESeq, edgeR and another recently introduced package, PoissonSeq, using simulated data.

Key Words: Gene expression, next generation sequencing, over-dispersed data.