Controlling False Discovery Rates in RNA-Sequencing Data

Conrad Burden[1], Sumaira Qureshi[1], and Susan Wilson[1,2]
[1] Australian National University, Canberra, ACT, Australia
[2] University of New South Wales, Sydney, NSW, Australia
[4] Corresponding author: Conrad Burden, e-mail: conrad.burden@anu.edu.au

Abstract

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.

Keywords: gene expression, next generation sequencing, over-dispersed data

1. Introduction

Transcriptome-wide expression profiling is accomplished from high throughput sequencing (HTS) technology via the technique of RNA-sequencing (RNA-seq) in which RNA transcripts sampled from a biological source are fragmented to convenient lengths, reverse transcribed to cDNA, amplified, sequenced and the reads identified by mapping to a reference genome. A summary of the RNA-seq procedure is given in the introductory material to Li et al. (2012). A number of software packages have been developed specifically for the purpose of analysing tables of read counts from biological replicate sequencing runs under two or more conditions with the specific purpose of detecting which genes are differentially expressed (DE) and quantifying the degree of differential expression via p-values and estimated false discovery rates (FDRs). An extensive comparison of the performance of eleven such packages has recently been published by Soneson and Delorenzi (2013).

HTS count data is well represented as over-dispersed Poisson data, as the Poisson shot noise inherent in sampling a relatively small number of reads from a large number of molecules in solution is compounded with biological variability and with technical variability due to sample preparation. Two of the most sophisticated packages for detecting differential expression from RNA-seq data, namely edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010), model the over-dispersed Poisson count data using a negative binomial (NB) model. The read counts for the biological replicates for each gene in each condition are fitted to a NB distribution via an algorithm that involves borrowing information from count data for

the complete set of genes. A transcript abundance for each gene is then inferred from the gene's NB mean. The null hypothesis corresponding to no differential expression is that the transcript abundance is the same in both conditions. Both packages provide p-values from which estimates of FDRs are extracted using the Benjamini-Hochberg algorithm (Benjamini and Hochberg, 1995). For concise summaries of differences between edgeR and DESeq see Robles et al. (2012) and the supplementary material to Soneson and Delorenzi (2013).

A recent addition by Li et al. (2012) to the suite of available packages for analysing RNA-seq data is PoissonSeq. This method power-transforms over-dispersed count data to (non-integer valued) quasi-Poisson data and normalises by iteratively determining a subset of genes satisfying a null-hypothesis Poisson model. This subset is typically chosen to be half the total number of genes and is interpreted as falling within the fraction π_0 of non-DE genes. An unsigned score statistic, which has a χ^2-distribution under the null hypothesis for the Poisson log-linear model described in Section 3 of Li et al. (2012), is used to detect differential expression. The FDR is estimated using a novel modified plug-in estimate in which the permutation distribution of the score statistic is calculated only from genes which are likely to be null. Using evidence of experiments with synthetic NB data, Li et al. claim that their method achieves considerably improved estimates of the FDR compared with edgeR.

This paper introduces an extension to the packages edgeR and DESeq which we call Polyfit. The aim of Polyfit is to improve calculations of p-values and estimates of the FDR by replacing the Benjamini-Hochberg procedure with an adapted version of a procedure for multiple hypothesis testing proposed by Storey and Tibshirani (2003). The secondary purpose of this paper is to perform a comparative analysis of the five packages PoissonSeq, edgeR, DESeq, and our extended versions Polyfit-edgeR and Polyfit-DESeq using synthetic data.

2. Methods

The packages edgeR and DESeq are state-of-the-art, but nevertheless are subject to shortcomings resulting from the computational complexity of estimating the parameters of the assumed NB distribution for each gene. To illustrate this, Fig. 1(a) shows the nominal p-value spectrum obtained from the DESeq algorithm for simulated data of 4 replicates of control and 4 replicates of treatment cases for 46,446 genes created with a range of means and over-dispersions typical of that found in the human transcriptome. In these data, the mean expression of 15% of the treatment genes have been up- or down-regulated by at least a factor of 2 relative to the control data. For the purposes of the current illustration, and as part of the implementation of our method, we have made changes to the original DESeq and edgeR algorithms to smooth out an artefact spike at $p = 1$ resulting from estimating p-values from a discrete distribution. Nevertheless, we observe that even with this spike redistributed the p-value spectrum for the 85% of genes which are unregulated (shaded) is far from uniform. The effect is more pronounced for DESeq than for edgeR. Using a false positive rate to control for differential expression with these calculated p-values for the null-hypothesis genes would lead to an overly conservative measure of significance and hence loss of power to detect differential expression.

DESeq and edgeR correct for multiple hypothesis testing via the Benjamini-Hochberg procedure. For each gene an 'adjusted p-value' (also known as q-value) is calculated to enable the expected false discovery rate (FDR) (i.e. the proportion of positives returned which are false positives) to be used to control for differential expression. Herein we propose an alternate method for estimating p-values and q-values by adapting an alternate procedure due to Storey and Tibshirani (2003). This is a graphical procedure for multiple hypothesis testing in which the proportion of cases satisfying the null hypothesis is estimated from the behaviour of the p-value spectrum as $p \to 1$, enabling estimates of q-values to be obtained graphically at any

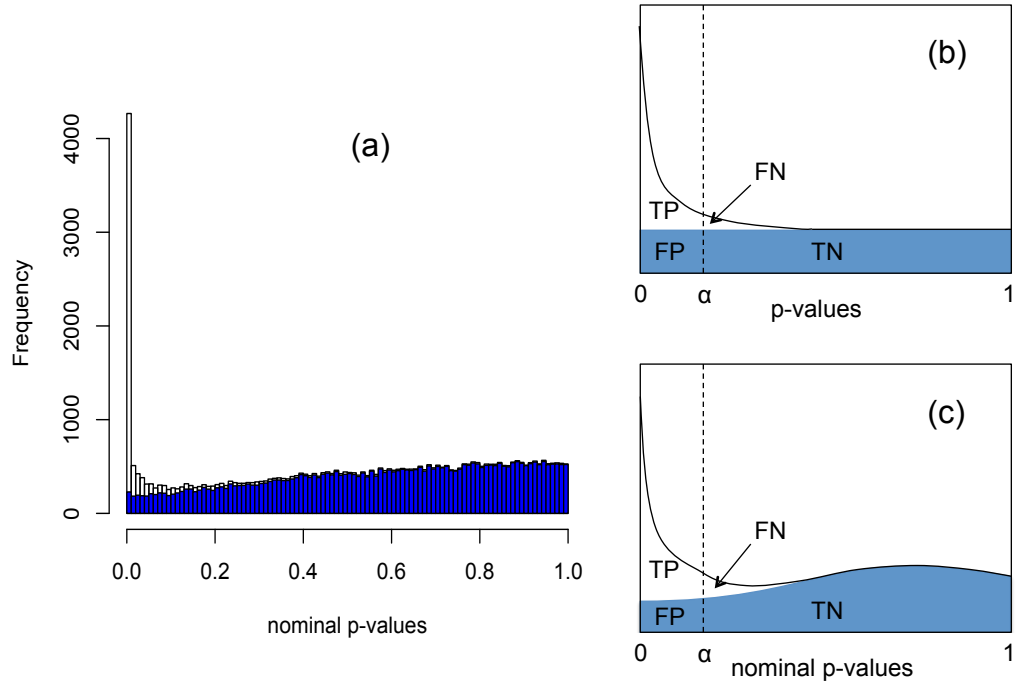

Figure 1: (a) Nominal p-value spectrum calculated by DESeq for synthetic data RNA-seq with 15% genes up- or down-regulated. The shaded histogram is the 85% of transcripts which are unregulated. (b) Schematic representation of the Storey-Tibshirani procedure for correcting for multiple hypothesis testing, assuming correctly calculated p-values. (c) Schematic representation of the Storey-Tibshirani procedure adapted to RNA-seq data. By 'nominal p-values' we mean p-values as calculated by a computer package relying on a NB model using estimated parameters, such as DESeq or edgeR. (TP = true positives, FP = false positives, FN = false negatives and TN = true negatives at a specified significance point α.)

p-value α as the ratio FP/(TP + FP) (see Fig. 1(b)). The procedure implicitly assumes p-values are calculated accurately and have a uniform distribution under the null hypothesis.

The adaptation of the Storey-Tibshirani procedure to RNA-seq data is illustrated in Fig. 1(c): A quadratic function is fitted to the right hand part of the nominal p-value spectrum supplied by the existing software and extrapolated to the complete interval [0, 1]. The area under the extrapolated curve is assumed to approximate the histogram of nominal p-values for the non-DE genes. Corrected p-values and q-values are then estimated at each nominal p-value α from the formulae

$$p_{\text{corrected}} = \text{FP}/(\text{FP} + \text{TN}), \qquad q_{\text{corrected}} = \text{FP}/(\text{FP} + \text{TP}).$$

The method provides an estimate of the proportion π_0 of genes satisfying the null hypothesis of no differential expression as the shaded area divided by the total number of genes, and hence also an estimate of the fraction $1 - \pi_0$ of DE genes. We refer to our adapted Storey-Tibshirani procedure, which we have implemented as a set of R functions, as 'Polyfit' (for polynomial fit). A detailed description of the algorithm and the source code will be published elsewhere (Burden et al., 2013).

3. Results

We tested the relative performance of PoissonSeq, the original DESeq and edgeR, and our extensions Polyfit-DESeq and Polyfit-edgeR using synthetic datasets. Each dataset consists of NB distributed counts simulating n replicates control data and n replicates

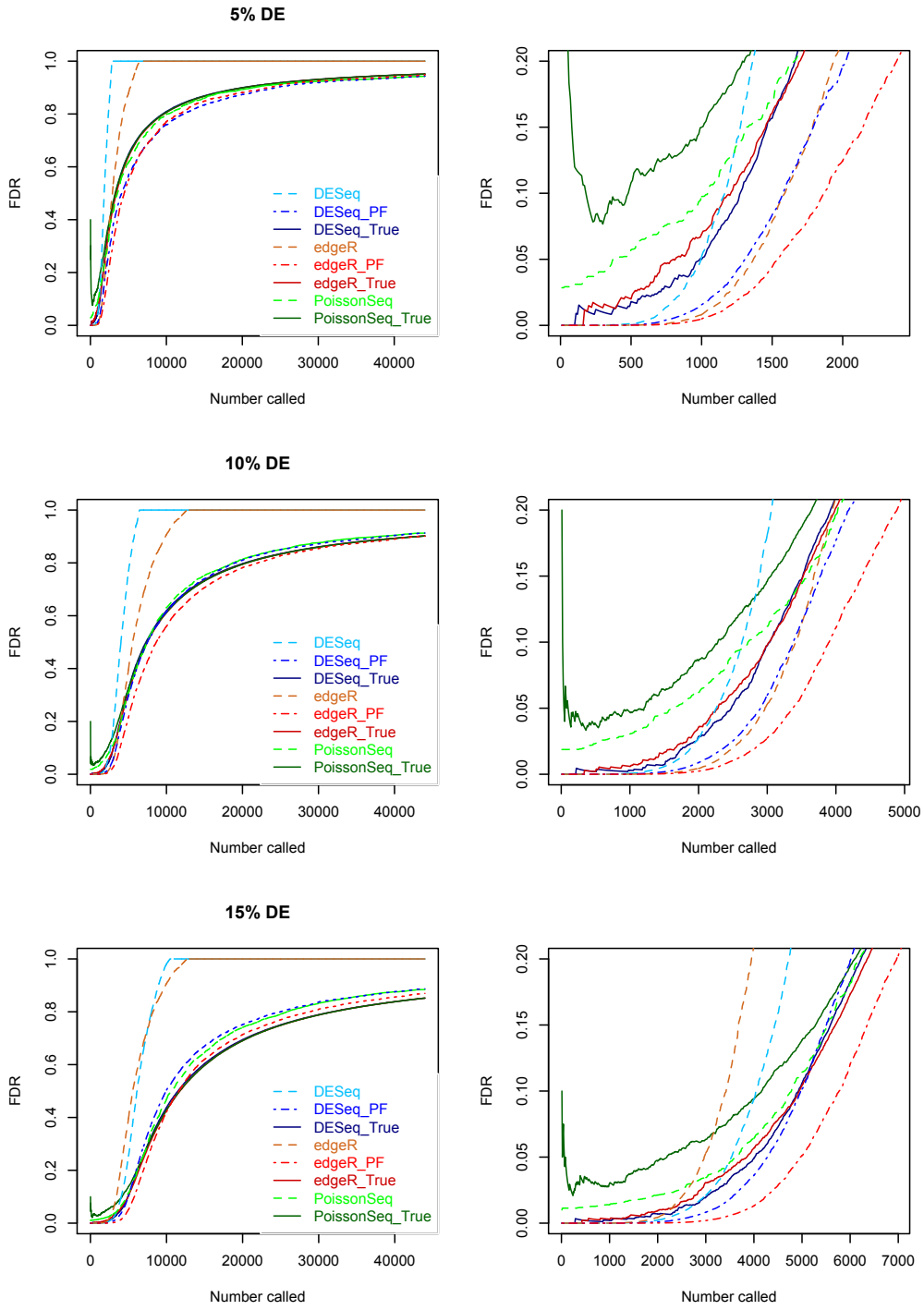

Figure 2: True (solid curves) and estimated (broken curves) FDRs for n = 4 control and 4 treatment replicates of synthetic data with 5, 10 and 15% DE genes. Five different methods are used for calculating p-values and q-values: PoissonSeq (Li et al., 2012), DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), and our proposed variants Polyfit-DESeq and Polyfit-edgeR (labelled with the extension PF)`. The true FDR curves do not differ noticeably on the scale of the plots between DESeq and Polyfit-DESeq or between edgeR and Polyfit-edgeR. The right hand plots are an expanded view of the subset of genes up to a significance point roughly corresponding to the number of truly DE genes.

of treatment data in which a specified percentage of genes are DE by at least a factor of 2. The calculation was done for n = 2, 4, 6 and 10 simulated biological replicates and 1, 5, 10 and 15% of a total of 46,446 genes DE.

The left-hand plots in Fig. 2 show true and estimated FDRs calculated from synthetic data over the complete range of p-values for the case n = 4. The plots confirm Li et al.'s (2012) findings that PoissonSeq substantially corrects an overestimation of the true FDR by the Benjamini-Hochberg procedure used by edgeR and DESeq as the significance point is raised to include a large number of genes called as being DE. The plots also show that this shortcoming of edgeR and DESeq is rectified by our adapted Storey-Tibshirani procedure, which brings Polyfit-edgeR and Polyfit-DESeq into close agreement with PoissonSeq and the true FDR curves. This observation holds in general provided at least 5% of the genes in the synthetic data are DE (Burden, 2013).

An issue not examined in the left-hand plots in Fig. 2 or in the simulations of Li et al. is the relative performance of different packages and methods for the subset of genes called as being most significantly DE. In the right hand plots of Fig. 2 we show the portion of the FDR curves up to a significance point roughly corresponding to the number of DE genes in each simulation. Out of a total of 46,446 genes, this corresponds to ~2,300, ~4,600 and ~7,000 genes for 5, 10 and 15% DE respectively. These plots indicate two disadvantages of PoissonSeq, namely that for the genes called as being most significantly DE, the true FDR is consistently higher than for the remaining four methods, and that the true FDR is under-reported by PoissonSeq. This is observed to occur out to a significance point corresponding to half the number of truly DE genes in all of our simulations (Burden, 2013), including those shown in the right hand plots of Fig. 2.

The ability of the remaining methods to estimate the FDR for the genes called as being most significantly DE varies according to the level of differential expression and the number of simulated biological replicates. Observations from FDR plots of our complete set of simulations (Burden, 2013) out to a significance point corresponding to half the number of truly DE genes are summarised in Fig. 3. At low levels of differential expression (up to 5% DE) or small numbers of simulated biological replicates ($n \leq 4$) all methods under-report the true FDR for the genes covered by Fig. 3. The Polyfit addition to edgeR and DESeq tends to lower the estimated FDR, thus exacerbating this problem. However, for higher levels of differential expression ($\geq$ 15% DE for edgeR and $\geq$ 10% DE for DESeq) and higher numbers of simulated biological replicates ($n \geq 6$ for edgeR and $n \geq 4$ for DESeq), DESeq and edgeR over-report the FDR over almost the whole range of genes. The Polyfit procedure attempts to correct this over-reporting, the effect of which is to give an accurate estimate of the true FDR for sufficiently high numbers of biological replicates ($n \geq 10$ for edgeR or $n \geq 6$ for DESeq). One would in any case recommend higher numbers of replicates via multiplexing as simulations with synthetic data demonstrate that this leads to considerable gains in power to detect DE (Robles et al., 2012).

4. Conclusions

We have introduced an add-on to the NB-based packages edgeR and DESeq for two-class detection of differential expression called Polyfit which achieves two of the advantages associated with the recently introduced package PoissonSeq: (1) it provides an estimate of the fraction π_0 of non-DE genes, and (2) it gives an accurate estimate of the FDR over most of the range of the p-value spectrum (Fig. 2, left hand plots). Of more immediate interest to practising biologists is the software's performance for the genes called as being most significant, that is, the few hundred or so genes with the lowest p-values (Fig. 2, right hand plots). Our simulations with synthetic data demonstrate that the Polyfit extension to edgeR or DESeq will give an accurate estimate of the true FDR over the complete range of significance points,

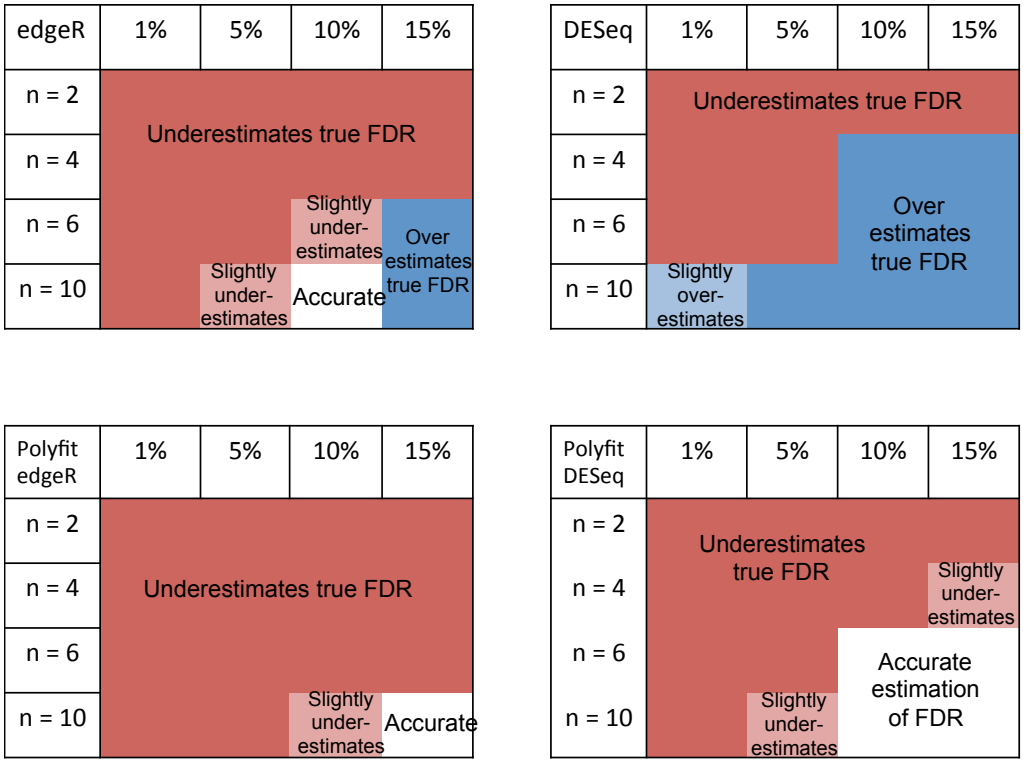

Figure 3: Summary of performance of the packages edgeR, DESeq and their Polyfit extensions in estimating the FDR for genes out to a significance point corresponding to half the number of truly DE genes. Experiments were done with n = 2,4,6 and 10 simulated replicates of synthetic data in which 1, 5, 10 and 15% of genes are DE.

including the subset of genes called as being most significantly DE, provided the number of replicates and percentage of differentially expressed genes is sufficiently high ($n \geq 10$ for edgeR or $n \geq 6$ for DESeq). This number of replicates is within the capabilities of current sequencing technologies by use of multiplexing in situations where budgets are limited.

References

Anders, S. and Huber, W. (2010) "Differential expression analysis for sequence count data", *Genome Biology*, 11, R106.

Li J., Witten D.M., Johnstone, I.M., and Tibshirani, R. (2012) "Normalization, testing, and false discovery rate estimation for RNA-sequencing data", *Biostatistics*, 13, 523-538.

Burden, C.J., Qureshi, S.E. and Wilson, S.R. (2013) "Improved error estimates for the analysis of differential expression from RNA-seq data", (in preparation).

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, 26, 139-140.

Robles, J.A., Qureshi, S., Stephen, S., Wilson, S.R., Burden, C.J. and Taylor, J.M. (2012) "Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing", *BMC Genomics*, 13, 484.

Soneson, C. and Delorenzi, M. (2013) "A comparison of methods for differential expression analysis of RNA-seq data", *BMC Bioinformatics*, 14, 91.

Storey, J.D. and Tibshirani, R. (2003) "Statistical significance for genomewide studies", *Proceedings of the National Academy of Science*, 100, 9440-9445.