Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson¹,²,*†, Davis J. McCarthy²† and Gordon K. Smyth²

¹Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and 
²Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 
Victoria 3052, Australia

Received on March 29, 2009; revised on October 19, 2009; accepted on October 23, 2009

Advance Access publication November 11, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

It is expected that emerging digital gene expression (DGE) technologies will overtake microarray technologies in the near future for many functional genomics applications. One of the fundamental data analysis tasks, especially for gene expression studies, involves determining whether there is evidence that counts for a transcript or exon are significantly different across experimental conditions. edgeR is a Bioconductor software package for examining differential expression of replicated count data. An overdispersed Poisson model is used to account for both biological and technical variability. Empirical Bayes methods are introduced to moderate the degree of overdispersion across transcripts, improving the reliability of inference. The methodology can be used even with the most minimal levels of replication, provided at least one phenotype or experimental condition is replicated. The software may have other applications beyond sequencing data, such as proteome peptide count data.

Availability: The package is freely available under the GPL licence from the Bioconductor web site (http://bioconductor.org).

Contact: mrobinson@wehi.edu.au

1 INTRODUCTION

Modern molecular biology data present major challenges for the statistical methods that are used to detect differential expression, such as the requirement of multiple testing procedures and increasingly, empirical Bayes or similar methods that share information across all observations to improve inference. For microarrays, the abundance of a particular transcript is measured as a fluorescence intensity, effectively a continuous response, whereas for digital gene expression (DGE) data the abundance is observed as counts.

This note describes the software package edgeR (empirical analysis of DGE in R), which forms part of the Bioconductor project (Gentelman et al., 2004). edgeR is designed for the analysis of replicated count-based expression data and is an implementation of methodology developed by Robinson and Smyth (2007, 2008). Although initially developed for serial analysis of gene expression (SAGE), the methods and software should be equally applicable to emerging technologies such as RNA-seq (Li et al., 2008; Marioni et al., 2008) giving rise to digital expression data. edgeR may also be useful in other experiments that generate counts, such as ChIP-seq, in proteomics experiments where spectral counts are used to summarize the peptide abundance (Wong et al., 2008), or in barcoding experiments where several species are counted (Andersson et al., 2008). The software is designed for finding changes between two or more groups when at least one of the groups has replicated measurements.

2 MODEL

Bioinformatics researchers have learned many things from the analysis of microarray data. For instance, power to detect differential expression can be improved and false discoveries reduced by sharing information across all probes. One such procedure is limma (Smyth, 2004), where an empirical Bayes model is used to moderate the probe-wise variances. The moderated variances replace the probe-wise variances in the t- and F-statistic calculations. In a closely analogous but mathematically more complex procedure, edgeR models count data using an overdispersed Poisson model, and uses an empirical Bayes procedure to moderate the degree of overdispersion across genes.

We assume the data can be summarized into a table of counts, with rows corresponding to genes (or tags or exons or transcripts) and columns to samples. For RNA-seq experiments, these may be counts at the exon, transcript or gene-level. We model the data as negative binomial (NB) distributed,

\[ Y_{gi} \sim \text{NB}(M_i p_{gj}, \phi_g) \]  

for gene \( g \) and sample \( i \). Here, \( M_i \) is the library size (total number of reads), \( \phi_g \) is the dispersion and \( p_{gj} \) is the relative abundance of gene \( g \) in experimental group \( j \) to which sample \( i \) belongs. We use the NB parameterization where the mean is \( \mu_{gi} = M_i p_{gj} \) and variance is \( \mu_{gi} (1 + \mu_{gi} \phi_g) \). For differential expression analysis, the parameters of interest are \( p_{gj} \).

The NB distribution reduces to Poisson when \( \phi_g = 0 \). In some DGE applications, technical variation can be treated as Poisson. In general, \( \phi_g \) represents the coefficient of variation of biological variation between the samples. In this way, our model is able to separate biological from technical variation.
The required inputs for edgeR are straightforward and have been implemented recently. At present, testing for differential expression is supported only for pairwise comparisons; the user must specify which two groups to compare. We are currently investigating tests for more general cases.

Many of the early RNA-seq datasets involve sequence reads from technical replicates (e.g. same source of RNA) as opposed to biological replicates (e.g. RNA from different individuals). Technical replicates will generally have lower variability than biological replicates and in our experience, the dispersion parameter (and the moderation procedure in edgeR) may not be necessary. For experiments with technical replicates, the data may be fitted well by the Poisson distribution, as demonstrated in Marioni et al. (2008).

Since the Poisson distribution is a special case of the NB distribution (\(\phi=0\)), edgeR can perform a Poisson-based analysis. The pairwise exact testing procedure will still be useful for these datasets.

### 4 DISCUSSION

We have developed a Bioconductor package edgeR that addresses one of the fundamental downstream data analysis tasks for count-based expression data: determining differential expression. The package and methods are general, and can work on other sources of count data, such as barcoding experiments and peptide counts. To the authors’ knowledge, edgeR is the only software for SAGE or DGE data at this time which can account for biological variability when there are only one or two replicate samples.

**Funding:** National Health and Medical Research Council Program (Grant 406657 to G.K.S.); NHMRC, Independent Research Institutes Infrastructure Support Scheme (Grant 361646); Victorian State Government OIS grant (awarded to the WEHI); a Melbourne International Research Scholarship (to M.D.R.); Belz, Harris and IBS Honours scholarships (to D.J.M.).

**Conflict of Interest:** none declared.

### REFERENCES


![MA plot using common dispersion](image)

Fig. 1. DGE data can be visualized as ‘MA’ plots (log ratio versus abundance), just as with microarray data where each dot represents a gene. This plot shows RNA-seq gene expression for DHT-stimulated versus Control LNCaP cells, as described in Li et al. (2008). The smear of points on the left side signifies that genes were observed in only one group of replicate samples and the points marked ‘x’ denote the top 500 differentially expressed genes.