Bioinformatics, 31(15), 2015, 2589–2590
doi: 10.1093/bioinformatics/btv209
Advance Access Publication Date: 21 April 2015
Applications Note

OXFORD

Gene expression

edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test

Emmanuel Dimont[1,*], Jiantao Shi[1], Rory Kirchner[1] and Winston Hide[1,2,3]

[1]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, MA 02115, USA, [2]Harvard Stem Cell Institute, 1350 Massachusetts Ave, Cambridge, MA 02138, USA and [3]Sheffield Institute of Translational Neuroscience, University of Sheffield, 385a Glossop Road, Sheffield, S10 2HQ, UK

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on September 18, 2014; revised on March 16, 2015; accepted on April 13, 2015

Abstract

Summary: Next-generation sequencing platforms for measuring digital expression such as RNA-Seq are displacing traditional microarray-based methods in biological experiments. The detection of differentially expressed genes between groups of biological conditions has led to the development of numerous bioinformatics tools, but so far, few exploit the expanded dynamic range afforded by the new technologies. We present edgeRun, an R package that implements an unconditional exact test that is a more powerful version of the exact test in edgeR. This increase in power is especially pronounced for experiments with as few as two replicates per condition, for genes with low total expression and with large biological coefficient of variation. In comparison with a panel of other tools, edgeRun consistently captures functionally similar differentially expressed genes.
Availability and implementation: The package is freely available under the MIT license from CRAN (http://cran.r-project.org/web/packages/edgeRun).
Contact: edimont@mail.harvard.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next generation sequencing technologies are steadily replacing microarray-based methods, for instance transcriptome capture with RNA-Seq (Mortazavi *et al.*, 2008) and CAGE-Seq capture for the promoterome (Kanamori-Katayama *et al.*, 2011). All of these approaches result in digital expression data, where reads or tags are sequenced, mapped to the genome and then counted. The discrete nature of the data has required the development of new bioinformatics tools for their analysis that address discrete count data.

Once the expression has been quantified, an important next step is the statistical significance testing of differential expression between two or more groups of conditions. By the far the simplest and most popular approach reduces differential expression to a pairwise comparison of mean parameters, resulting in a fold-change measure of change and a *P*-value to ascertain statistical significance of the finding. To address this problem, tools such as *edgeR* (Robinson *et al.*, 2010), *DESeq2* (Love *et al.*, 2014) among many others have been developed and can be applied to any experiment in which digital count data is produced.

This vast array of tool choices can be bewildering for the biologist since it is generally not clear under which conditions a tool is more appropriate than its alternates. Traditional metrics used when benchmarking methods such as the false positive rate and power are useful but limited as they are purely statistical concepts that can only be tested on simulated data. Moreover, they do not help in determining to what extent methods deliver truly biologically important genes. This is a major challenge because in the vast majority of cases, we do not know what the true positives and negatives are.

In this article, we propose a novel metric to determine the number of functionally relevant genes reported by a differential expression

tool and present *edgeRun*, an extension of the *edgeR* package delivering increased power to detect true positive differences between conditions without sacrificing on the false positive rate. We show using simulations and a real data example that *edgeRun* is uniformly more powerful than a host of differential expression tools for small sample sizes. We also demonstrate how even though it may be less statistically powerful than *DESeq2* in some simulation cases, *edgeRun* nonetheless produces results that are functionally more relevant.

2 Methods

2.1 edgeRun: exact unconditional testing

Assuming independent samples, Robinson *et al.* (2010) proposed *edgeR*, an R package that eliminates the nuisance mean expression parameter by conditioning on a sufficient statistic for the mean, a strategy first popularized by Fisher (1925) for the binomial distribution. This leads to a calculation of the exact *P*-value that does not involve the mean. The advantage of this approach is its analytic simplicity and fast computation, however, a key disadvantage is that this conditioning approach loses power, especially for genes whose counts are small.

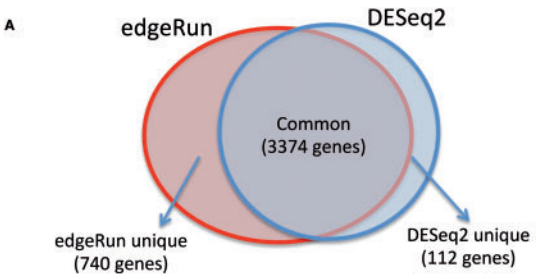
We propose an alternative more powerful approach which eliminates the nuisance mean parameter via maximizing the exact *P*-value over all possible values for the mean without conditioning which we call 'unconditional *edgeR*' or *edgeRun*. This technique was initially proposed by Barnard (1945) for the binomial distribution. The main disadvantage of this method is the higher computational burden required for the maximization step. On the other hand, the gain in power can be significant. A thorough derivation and comparison of both methods can be found in the Supplementary Methods.

2.2 Benchmarking against other methods

The *compcodeR* Bioconductor package (Soneson, 2014) was used to benchmark the performance of *edgeRun* against a panel of available other tools using a combination of simulated and real datasets. *edgeRun* had the highest area under the curve (AUC) of all methods and it maintained a comparable false discovery rate (FDR) similar to other tools. In terms of power, only *DESeq2* was found to outperform *edgeRun*. For this reason in the next section, we perform a functional comparison only with *DESeq2*. The full results are summarized in Supplementary Methods.

2.3 Comparing functional relevance

We propose to compare the genes called significant by various differential expression tools. Figure 1 compares the results of *edgeRun* and *DESeq2* applied to a prostate cancer dataset (Li *et al.*, 2008) using an FDR < 5% cutoff. Out of the 4226 genes reported as differentially expressed, 80% were common to both tools. The highest 500 up- or down-regulated of these consensus genes by fold-change are used as a seed signature. It is reasonable to hypothesize that true differentially expressed genes uniquely reported by a differential expression tool are functionally connected to genes in the consensus group. We use GRAIL (Raychaudhuri *et al.*, 2009) coupled with a global co-expression network COXPRESdb (Obayashi *et al.*, 2013) to assess the relatedness between a gene and the consensus group. As expected, nearly half of these seed genes are correlated with other members of the seed group, meaning that these consensus genes form a tightly connected network. Figure 1 shows that *edgeRun* reports 6.6 times more unique differentially expressed genes, and a larger proportion of which are co-expressed with the consensus: 33% of genes unique to *edgeRun* as compared with 17% of genes unique to *DESeq2* (*P*-value < 0.001). This means that the genes reported by *edgeRun* are more likely to be functionally relevant as they are more correlated with the consensus network. More details on this approach on evaluating functional relevance can be found in the Supplementary Methods.



B

	# Genes	# / % genes correlated to UP seeds	# / % genes correlated to DN seeds
Common UP	500	237 / 47.4%	-
Common DN	500	-	217 / 43.4%
edgeRun Unique	740	**109 / 14.7%**	**137 / 18.5%**
DESeq2 Unique	112	12 / 10.7%	7 / 6.2%

Fig. 1. Comparing the functional relevance of genes called significantly differentially expressed by edgeRun and DESeq2

3 Discussion

We present *edgeRun*, an R package that improves on the popular package *edgeR* for differential digital expression by providing the capability to perform unconditional testing, resulting in more power to detect true differences in expression between two biological conditions. Even though the computational burden is increased, the power gained using this approach is significant, allowing researchers to detect more true positives, especially for cases with as few as two replicates per condition and for genes with low expression, all the while without sacrificing on type-I error rate control. *edgeRun* is simple to use, especially for users already experienced with *edgeR* as it is designed to interface with *edgeR* objects directly, taking inputs and generating output in the same format.

Acknowledgements

We would like to thank Oliver Hofmann, Shannan Ho Sui, Gabriel Altschuler and Yered Pita Juarez for their valuable feedback.

Conflict of Interest: none declared.

References

Barnard,G.A. (1945) A new test for 2x2 tables. *Nature*, **156**, 177.

Fisher,R.A. (1925) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Kanamori-Katayama,M. *et al.* (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.*, **21**, 1150–1159.

Li,H. *et al.* (2008) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *PNAS*, **105**, 20179–20184.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, **15**, 550

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Obayashi,T. *et al.* (2013) COXPRESSdb: a database of comparative gene coexpression networks of eleven species of mammals. *Nucleic Acids Res.*, **41**, D1014–D1020.

Raychaudhuri,S. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Soneson,C. (2014) compcodeR-an R package for benchmarking differential expression methods for RNA-Seq data. *Bioinformatics*, **30**, 2517–2518.