A survey of statistical software for analysing RNA-seq data

Dexiang Gao,[1,5] Jihye Kim,[2] Hyunmin Kim,[4] Tzu L. Phang,[3] Heather Selby,[2] Aik Choon Tan[2,5] and Tiejun Tong[6]***

[1]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA
[2]Division of Medical Oncology, University of Colorado School of Medicine, Aurora, CO 80045, USA
[3]Division of Critical Care and Pulmonary Medicine, University of Colorado School of Medicine, Aurora, CO 80045, USA
[4]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA
[5]Department of Biostatistics and Informatics, University of Colorado School of Public Health, Aurora, CO 80045, USA
[6]Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA
**Correspondence to:* Tel: +1 303 724 4356; Fax: +1 303 724 4491; E-mail: Dexiang.Gao@UCDenver.edu
**Tel: +852 3411 7340; Fax: +852 3411 5811; E-mail: tongt@hkbu.edu.hk

Date received (in revised form): 26th May 2010

Abstract

High-throughput RNA sequencing is rapidly emerging as a favourite method for gene expression studies. We review three software packages — edgeR, DEGseq and baySeq — from Bioconductor http://bioconductor.org for analysing RNA-sequencing data. We focus on three aspects: normalisation, statistical models and the testing employed on these methods. We also discuss the advantages and limitations of these software packages.

Keywords: *statistical software, RNA-sequencing analysis, normalisation, sequencing data*

Introduction

High-throughput genome-wide RNA profiling by deep sequencing (RNA-seq) is rapidly emerging as a favourite method for gene expression studies. RNA-seq provides more precise measurement of levels of transcripts at a wide dynamic range and the ability to quantitate and detect known and novel isoforms by comparison with hybridisation-based technology (oligonucleotide and cDNA microarrays). In every sequencing run, tens of millions of short reads are simultaneously sequenced in each lane by the next generation sequencer. After pre-processing and mapping against a reference genome, the total number of counts for each mappable transcript is reported. It has been reported that the sequencing results are highly reproducible.[1] One of the main applications of RNA-seq is to identify differential expression (DE) genes under two or more different phenotypes (eg cancer versus normal samples).

Several statistical methods have been proposed to identify DE.[1–5] When choosing a statistical analysis approach, some aspects need to be considered:

(a) *Normalisation.* It was noticed that the observed number of reads for a gene depends on the expression level and the length of the gene, and also on the RNA composition of the sample.[6,7] The purpose of the normalisation is to minimise the influences of gene length and total sample RNA composition so that the normalised read counts represent a direct reflection of the targeted gene expression level. It has been shown that the normalisation procedure has a great impact on DE detection.[2,7] Depending on the experimental design, different normalisation methods are required.

(b) *Statistical model.* The Poisson distribution is commonly used to model count data. Due to biological and genetic variations, however, for sequencing data the variance of a read is often much greater than the mean value. That is, the data are over-dispersed. In such cases, one natural alternative to Poisson is the negative binomial (NB) model. In addition to these two commonly used models, other choices have also been proposed in the literature.[8,9]

(c) *Testing.* In terms of tag detection, there are mainly two types of methods: exact testing methods — such as Fisher's exact test (FET), and tests based on large sample approximation.[1,10,11]

In this paper, we review three publicly available software packages from Bioconductor, which are specifically designed for RNA-seq data analyses. Our main goal was to provide detailed descriptions for each package to guide software selections for identifying DE for a given study design.

Software packages surveyed

1. edgeR

The R Bioconductor package, edgeR,[12] provides statistical routines for detecting DE in RNA-seq data. The package is extremely flexible and can handle the count data irrespective of whether or not they are over-dispersed. If the data are over-dispersed, the NB model is used. Conversely, the Poisson model is used when there is no over-dispersion detected in the data. edgeR requires the data to be in either one of two formats: a single file containing a table of counts, with the first column containing the read (refer to 'tag' in the package) identifiers and the remaining columns containing the tag counts for each sample sequenced; or an individual file for each library, each with the first column for tag identifier and second column for counts.

Normalisation

The quantile-adjusted method is used to standardise total read counts (library sizes) across samples.[11] Samples are assumed to be independent and identically distributed from NB distribution (M*p, ϕ) (see details in the Model section), where ϕ is initially estimated from all the samples, M* is the geometric mean of original library sizes and p, the proportion of tag g in the sequenced sample, can then be estimated, providing values for M and ϕ. Linear interpolation of quantile function is used to equate the quantiles across samples. The process is updated with new ϕ and p until ϕ converges.

Model

edgeR is based on NB distribution. Let Y_{gij} denote the observed data; where g is the gene (tag, exon, etc.), i is the experimental group and j is the index of samples. The counts can be modelled as $Y_{gij} \sim NB(M_j p_{gi,} \phi_g)$, with mean $\mu_{gi} = M_j p_{gi}$ and variance $= \mu_{gi} + \mu_{gi}^2 \phi$, where M_j represents the library size, (ie the sum of the counts of tags in a sample) and p_{gi} represents the proportion of tag g of the sequenced sample for group i. ϕ_g is the over-dispersion parameter (relative to Poisson) for accounting for biological or sample-to-sample variation. There are two options for ϕ_g in the package; one is to use a common dispersion for all the tags and the other is to assume tagwise dispersion.[3,11] For many applications, using a common dispersion will be adequate. For the tagwise dispersion, edgeR moderates the estimates towards a common dispersion. Moderation is determined using an empirical Bayes rule.[3] It is noted in general that tagwise dispersion penalises tags with great variability within groups. If the common dispersion estimate is much greater than 0, it indicates that there is more variability in the data than the Poisson model can account for, and NB distribution should be used. With $\phi_g = 0$, the NB distribution reduces to Poisson distribution.

Testing

edgeR employs an exact test for the NB distribution based on the normalised data. The test parallels with FET. The 'exactTest' function allows pairwise comparisons of groups. One of the objects produced by the function includes logFC, the log-fold change difference in the counts between the groups, and exact p-values. The results of the NB test can be accessed using the 'topTags' function, in which the adjusted *p*-values for multiple testing are reported using Benjamini and Hochberg's approach[13] as the default method of adjustment. Users can also supply their own desired adjustment method.

2. DEGseq

DEGseq is another R package specifically designed to identify DE from RNA-seq data.[9] The package includes two novel methods, the MA plot-based method (where M is the log ratio of the counts between two experimental conditions for gene *g*, and A is the two group average of the log concentrations of the gene) with a random sampling model (MARS) and the MA plot-based method with technical replicates (see details in the Models section), along with three existing methods: FET, the likelihood ratio test (LRT) and samWrapper. samWrapper was developed previously for microarray data analysis.[14] In this paper we focus our attention on MARS, FET and the LRT. Unlike edgeR, which allows over-dispersion, DEGseq assumes a binomial or Poisson distribution (which limits its application to data with no over-dispersion) and is extremely easy to use. The user needs only to specify the model, the normalisation methods and the input data, and the results will be saved to the user's designed folder. The input of the package is uniquely mapped reads (or tags), gene annotation of the corresponding genome and gene expression counts for each sample. The output includes a text file and a summary. The text file contains the original sample counts, *p*-value, and two *q*-values, indicating the expression difference between the two treatment groups for each gene, which are the adjusted *p*-values.

Normalisation

There are several choices for normalisation: 'none', 'loess' and 'median'. The recommended (or default) method is 'none.'

Models

(i) MARS

In the MARS model, RNA sequencing is modelled as a random sampling process.[15] Each read is sampled independently and uniformly from every possible nucleotide in the sample. The number of tags coming from a gene follows a binomial distribution, which can be approximated by a Poisson distribution. The model identifies and visualises DE genes based on the MA plot. It follows that M and A are both normally distributed, given that the samples from the two conditions are independent. The conditional distribution of M, given that A = a, is also normally distributed. Under the null hypothesis that the probabilities of the tag coming from a specific gene are the same between the two experimental conditions, a Z-score statistic for the gene can be calculated, and the *p*-value can be converted to indicate if the gene *g* is differentially expressed. The MA plot has been widely used to detect and visualise the intensity-dependent ratio of microarray data.[16]

(ii) LRT

LRT was used by Marioni[1] to identify differentially expressed genes from sequencing data. In the dataset used, samples from each group are technical replicates. Let Y_{gij} represent the number of reads mapped to gene *g* for the *j*-th lane (sample) from group *i*, as described earlier for the edgeR model. Y_{gij} then can be modelled as a Poisson random variable with mean $\mu_{gi} = M_j p_{gi}$. Under the null hypothesis, the two groups A and B have the same value for gene *g*, $p_{gi} = p_j$ and, under the alternative hypothesis, $p_{gi} = p_j^A$ for samples from group A and $p_{gi} = p_j^B$ for samples from group B. Poisson regression is then performed where the standard LRT is computed to test for differences in expression between the two groups.

(iii) FET

Under the random sample process, the number of reads from a gene follows a binomial distribution which can be approximated by a Poisson distribution. The group counts and the total counts are also Poisson distributed. FET is then used to calculate the probability of observing group counts as the observed or more extreme if the null hypothesis is true (no difference between the two groups in the expression of gene *g*) for each gene.

Testing

FET and large sample approximation, such as the LRT and Z score test, are the choices. Multiple testing was adjusted using the methods of either Benjamini and Hochberg[13] or Storey and Tibshirani.[17]

3. baySeq

baySeq differs from the above two packages by employing an empirical Bayesian analysis approach to determine if there is DE between two different conditions.[18] It begins by assuming that the data follow a distribution, either Poisson or NB, which is defined by a set of underlying parameters. The prior for each parameter is estimated by first bootstrapping from the data, and then either applying the maximum likelihood method (assuming that the prior is from a Poisson or NB distribution), or applying quasi-likelihood methods. For each gene or tag, two scenarios (hypotheses) are envisaged: one where the expression pattern is the same across the two conditions (ie that there is no DE between the two conditions for the gene); the other where the expression patterns differ between the two conditions (ie that there is DE for the gene). Given the prior estimates and the likelihood of the distribution of the data, one can estimate the posterior likelihood under the two scenarios to determine if there is DE for that gene or tag. Two distributions are proposed for the data; one is Poisson and the other is NB. The Poisson model is faster, yet the NB model provides better fit for most RNA-seq data. baySeq recommends using the NB model in general. The required data format is the same as the edgeR package. Parallel processing is provided through the 'snow' package for faster processing.

Normalisation

No normalisation procedure is proposed in this package.

Models

Two models are used in the package; one is to assume a Poisson distribution on each tag that is $Y_{gij} \sim (M_j p_{gi,})$, where the prior for p_{gi} is assumed to follow gamma distribution $p_{gi} \sim \Gamma(\alpha_{gi},\ \beta_{gi})$. This model is therefore named the Poisson-gamma approach. In general, a subset of data is taken initially and more than 5,000 iterations are recommended for the bootstrapping. Gamma parameters are calculated using maximum likelihood methods. The mean of the maximum likelihood estimates is taken to obtain a prior on p_{gi}. An initial prior value needs to be provided for the program to start. The other model assumes that the data are NB distributed, $Y_{gij} \sim NB(M_j p_{gi,}\ \phi_g)$. As there is no conjugate prior available for this distribution, a numerical solution for an empirical prior is required. The program first bootstraps from the data, with around 10,000 iterations suggested. The parameters for an empirical prior distribution are estimated using the quasi-likelihood approach. Given the prior and the likelihood of the data, the posterior likelihoods are then calculated. The program repeatedly bootstraps to improve the accuracy of the prior estimation, and the posterior likelihood is updated accordingly.

Testing

The estimated posterior likelihoods are reported on the natural logarithmic scale.

Conclusions

Among the three software packages surveyed, DEGseq is the easiest to use. baySeq in general takes much longer to run with the recommended number of iterations for the bootstrap. edgeR is the most flexible package and can handle both Poisson data and over-dispersed data without the need to pre-specify the model. baySeq also includes these two models but one needs to pre-specify which to use. DEGseq does not handle over-dispersed data. Over-dispersion is extremely common among biological samples. edgeR provides estimates of the over-dispersion parameter, which can be helpful in determining if a Poisson model is appropriate when applying the other two packages.

edgeR normalises the data by scaling the number of reads to a common value across all samples. Recent studies have shown that gene length and RNA composition also bias total read counts of targeted genes. New software, providing normalisation for gene length and RNA composition, will be expected in the future.

References

1. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. *et al.* (2008), 'RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays', *Genome Res*. Vol. 18, pp. 1509–1517.
2. Bullard, J.H., Purdom, E.A., Hansen, K.D., Durinck, S. *et al.* (2010), 'Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments', *BMC Bioinformatics* Vol. 11, p. 94.
3. Robinson, M.D. and Smyth, G.K. (2008), 'Small-sample estimation of negative binomial dispersion, with applications to SAGE data', *Biostatistics* Vol. 9, pp. 321–332.
4. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. *et al.* (2008), 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nat. Methods* Vol. 5, pp. 621–628.
5. Vencio, R.Z., Brentani, H., Patrao, D.F. and Pereira, C.A. (2004), 'Bayesian model accounting for within-class biological variability in serial analysis of gene expression (SAGE)', *BMC Bioinformatics* Vol. 5, p. 119.
6. Oshlack, A. and Wakefield, M.J. (2009), 'Transcript length bias in RNA-seq data confounds system biology', *Biol. Direct* Vol. 4, p. 14.
7. Robinson, M.D. and Oshlack, A. (2010), 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biol*. Vol. 11, p. R25.
8. Baggerley, K.A., Deng, L., Morris, J.S. and Aldaz, C.M. (2004), 'Overdispersed logistic regression for sage: Modelling multiple groups and covariates', *BMC Bioinformatics* Vol. 5, p. 144.
9. Wang, L., Feng, Z., Wang, X., Wang, X. *et al.* (2010), 'DEGseq: An R package for identifying differentially expressed genes from RNA-seq data', *Bioinformatics* Vol. 26, pp. 136–138.
10. Kal, A.J., Van Zonneveld, A.J., Benes, V., van den Berg, M. *et al.* (1999), 'Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources', *Mol. Biol. Cell* Vol. 10, pp. 1859–1872.
11. Robinson, M.D. and Smyth, G.K. (2007), 'Moderated statistical tests for assessing differences in tag abundance', *Bioinformatics* Vol. 23, pp. 2881–2887.
12. 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* Vol. 26, pp. 139–140.
13. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *J. R. Stat. Soc. Ser. B* Vol. 57, pp. 289–300.
14. Tusher, V.G., Tibshirani, R. and Chu, G. (2001), 'Significance analysis of microarrays applied to the ionizing radiation response', *Proc. Natl. Acad. Sci. USA* Vol. 98, pp. 5116–5121.
15. Jiang, H. and Wong, W.H. (2009), 'Statistical inferences for isoform expression in RNASeq', *Bioinformatics* Vol. 25, pp. 1026–1032.
16. Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M. *et al.* (2002), 'Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation', *Nucleic Acids Res*. Vol. 30: p. e15.
17. Storey, J. and Tibshirani, R. (2003), 'Statistical significance for genome-wide studies', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 9440–9445.
18. Hardcastle, T.J. and Kelly, K. (2010), 'Identifying patterns of differential expression in count data', *BMC Bioinformatics* Vol. 11, p. 422.