DEGseq: an R package for identifying differentially expressed genes from RNA-seq data

Likun Wang1,2, Zhixing Feng1, Xi Wang1, Xiaowo Wang1,* and Xuegong Zhang1,*
1MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.
2College of Computer Science and Technology, Jilin University, Changchun 130012, China.

ABSTRACT
Summary: High-throughput RNA sequencing (RNA-seq) is rapidly emerging as a major quantitative transcriptome profiling platform. Here we present DEGseq, an R package to identify differentially expressed genes or isoforms for RNA-seq data from different samples. In this package, we integrated three existing methods, and introduced two novel methods based on MA-plot to detect and visualize gene expression difference.
Availability: The R package and a quick-start vignette is available at http://bioinfo.au.tsinghua.edu.cn/software/degsseq
Contact: XW: xwwang@tsinghua.edu.cn; XZ: zhangxg@tsinghua.edu.cn.
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
High-throughput sequencing technologies developed rapidly in recent years. These technologies can generate millions of reads in a relatively short time and at low cost. Using such platforms to sequence cDNA samples (RNA-seq) has been shown as a powerful method to analyze the transcriptome of eukaryotic genomes (Wang et al., 2009). RNA-seq can provide digital gene expression measurement and is regarded as an attractive approach competing to replace microarrays for analyzing transcriptome in an unbiased and comprehensive manner.

Up to now, there are few handy programs for comparing RNA-seq data and identifying differentially expressed genes from the data, although some recent publications have described their methods for this task (Marioni et al., 2008; Bloom et al., 2009; Tang et al., 2009). Here, we present DEGseq, a free R package for this purpose. Two novel methods along with three existing methods have been integrated into DEGseq to identify differentially expressed genes. The input of DEGseq includes a data file and an XHTML summary page. The text file contains the expression values for the samples, a P-value and two kinds of Q-values for each gene to denote its expression difference between libraries. The XHTML summary page contains statistic summary report graphs as shown in Figure 1A.

2 METHODS
RNA sequencing could be modeled as a random sampling process, in which each read is sampled independently and uniformly from every possible nucleotides in the sample (Jiang et al., 2009). Under this assumption the number of reads coming from a gene (or transcript isoform) follows a binomial distribution (and could be approximated by a Poisson distribution). Based on this statistical model, Fisher’s exact test and likelihood ratio test were proposed to identify differentially expressed genes (Marioni et al., 2008; Bloom et al., 2009). The two methods have been integrated into DEGseq.

2.1 MA-plot-based method with random sampling model
Using the statistical model described above, we proposed a novel method based on the MA-plot, which is a statistical analysis tool having been widely used to detect and visualize intensity-dependent ratio of microarray data (Yang, et al., 2002). Let \( C_1 \) and \( C_2 \) denote the counts of reads mapped to a specific gene obtained from two samples, with \( C_i \sim \text{binomial}(n_i, p) \), \( i = 1, 2 \), where \( n_i \) denotes the total number of mapped reads and \( p \) denotes the probability of a read coming from that gene. We define \( M = \log_2 C_1 - \log_2 C_2 \), and \( A = (\log_2 C_1 + \log_2 C_2)/2 \). It can be proven that under the random sampling assumption the conditional distribution of \( M \) given that \( A = a \) (a is an observation of \( A \)), follows an approximate normal distribution (see Supplemental Methods Section 1). For each gene on the MA-plot, we do the hypothesis test of \( H_0: p_1 = p_2 \) versus \( H_1: p_1 \neq p_2 \). Then a P-value could be assigned based on the conditional normal distribution (see Supplemental Materials for detail).

2.2 MA-plot-based method with technical replicates
Though it has been reported that sequencing platform has low background noise (Marioni et al., 2008; Wang et al., 2009), technical replicates would still be informative for quality control and to estimate the variation due to different machines or platforms. We proposed another MA-plot-based method which estimates the noise level by comparing technical replicates in the data (if available). In this method, a sliding-window is first applied on the MA-plot of the two technical replicates along the \( A \)-axis to estimate the random variation corresponding to different expression levels. A smoothed estimate of the intensity-dependent noise level is done by loess
regression, and converted to local standard deviations of $M$ conditioned on $A$, under the assumption of normal distribution. The local standard deviations are then used to identify the difference of the gene expression between the two samples (see Supplemental Materials for detail).

2.3 Multiple testing correction
For the above methods, the $P$-values calculated for each gene are adjusted to $Q$-values for multiple testing corrections by two alternative strategies (Benjamini et al. 1995; Storey et al. 2003). Users can set either a $P$-value or a false discovery rate (FDR) threshold to identify differentially expressed genes.

2.4 Dealing with two groups of samples
To compare two sets of samples with multiple replicates or two groups of samples from different individuals (e.g. disease samples vs. control samples), we employed the R package samr (Tibshirani et al. 2009) in DEGseq. The package samr implemented the method described in Tusher et al. (2001) which assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements and uses permutations of the repeated measurements to estimate false discovery rate.

3 APPLICATION EXAMPLES
We applied DEGseq on the RNA-seq data from Marioni et al. (2008). The RNA samples from human liver and kidney were analyzed using the Illumina Genome Analyzer sequencing platform. Each sample was sequenced in seven lanes, split across two runs of the machine, and two different cDNA concentrations (1.5 pM and 3pM) were tested for each sample. We used the refFlat gene annotation file downloaded from UCSC Genome browser and chose the method proposed by Storey et al. (2003) to correct $P$-values for multiple testing.

We first checked whether the variation between technical replicates could be explained by the random sampling model. This was done with the “checking” feature in DEGseq (see Supplemental Material) on kidney sample sets kidneyR1L1 (sequenced in Run 1, Lane 1) and kidneyR1L3 which were generated at same cDNA concentration. Figure 1B shows that the variation can be almost fully explained by the random sampling model, which supports the notion that technical replicates of this data set have little technical variation (Marioni et al., 2008). And none gene was falsely identified as differentially expressed between the two replicates by each method at a FDR of 0.1% respectively (see Supplemental Table 1). However, samples sequenced at different concentrations showed larger variance (see Supplemental Fig. S1A).

We next applied DEGseq to compare the samples from kidney (kidneyR1L1) and liver (liverR1L2). For the MA-plot-based method that needs technical replicates, we used kidneyR1L1 and kidneyR1L3. More than 6000 genes were identified as differentially expressed by each method at a FDR of 0.1% respectively. And the lists of differentially expressed genes given by different methods are quite consistent with each other (see Supplemental Table S2). Figure 1C and Figure 1D shows the results given by the MA-plot based method with random sampling model and with technical replicates, respectively. And Supplemental Figure S1 shows the results given by the likelihood ratio test and Fisher’s exact test.

4 DISCUSSION
In some application, researchers may have several replicates sequenced under each condition. Current observations suggest that typically RNA-seq experiments have low technical background noise (which could be checked using DEGseq) and the Poisson model fits data well. In such cases, users could directly pool the technical replicates together to get higher sequencing depth and detect subtle gene expression changes. Otherwise the methods that estimate the noise by comparing the replicates are recommended. DEGseq also supports users to export gene expression values in a table format which could be directly processed by edgeR (Robinson 2009), an R package implementing the method based on negative binomial distribution to model overdispersion relative to Poisson for digital gene expression data with small replicates (Robinson et al. 2007).
DEGseq supports using expression values based on either the raw reads counts or normalized gene expression values like RPKM (Mortazavi et al., 2008). But for the methods based on the random sampling model, we suggest using the raw counts, which better fits the random sampling model.

DEGseq can also be applied to identify differential expression of exons or pieces of transcripts. Users can define their own "genes" and compare the expression difference of these "genes" using DEGseq by simply providing their own annotation files in UCSC refFlat format.

ACKNOWLEDGEMENTS

Funding: This work is supported in part by National Natural Science Foundation of China (grant numbers 30625012, 60721003, 60905013), the National Basic Research Program of China (2004CB518605) and Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University of China.

REFERENCES