BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 136–138
doi:10.1093/bioinformatics/btp612

Gene expression

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

Likun Wang[1,2], Zhixing Feng[1], Xi Wang[1], Xiaowo Wang[1,*] and Xuegong Zhang[1,*]

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua, University, Beijing 100084 and [2]College of Computer Science and Technology, Jilin University, Changchun 130012, China

Received on July 19, 2009; revised on September 30, 2009; accepted on October 19, 2009

Advance Access publication October 24, 2009

Associate Editor: Ivo Hofacker

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/degseq

Contact: xwwang@tsinghua.edu.cn; 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 (Bloom *et al.*, 2009; Marioni *et al.*, 2008; 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 is uniquely mapped reads from RNA-seq data with a gene annotation of the corresponding genome, or gene (or transcript isoform) expression values provided by other programs like RPKM (Mortazavi *et al.*, 2008). The output of DEGseq includes a text 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.

*To whom correspondence should be addressed.

2 METHODS

RNA sequencing could be modeled as a random sampling process, in which each read is sampled independently and uniformly from every possible nucleotide in the sample (Jiang and Wong, 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 (Bloom *et al.*, 2009; Marioni *et al.*, 2008). 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 $C_i \sim$ binomial (n_i, p_i), $i = 1,2$, where n_i denotes the total number of mapped reads and p_i 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 proved 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 Supplementary 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 Supplementary 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 (SDs) of M conditioned on A, under the assumption of normal distribution. The local SDs are then used

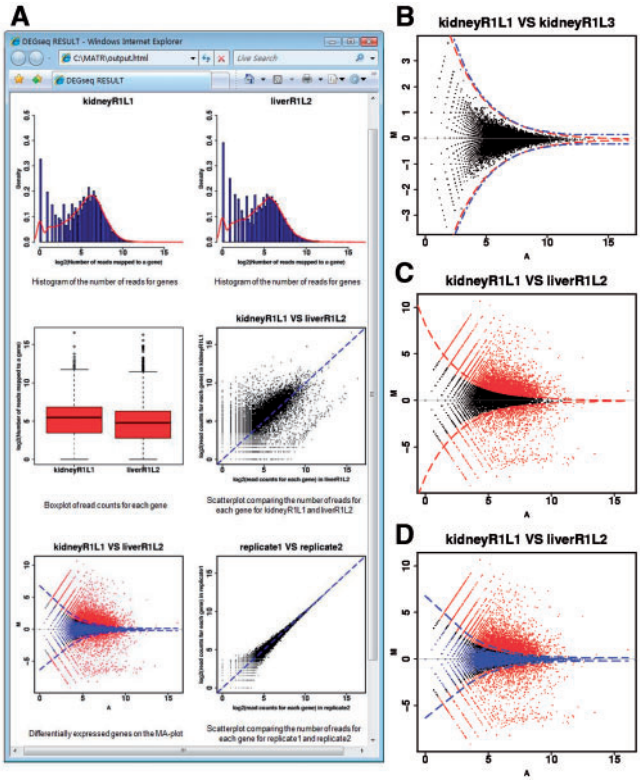

Fig. 1. (**A**) An example of the summary report page generated by DEGseq. (**B**) The plot generated by DEGseq showing whether the variation between technical replicates can be largely explained by the random sampling model. The red lines correspond to the 'theoretical' 4-fold local SD of M conditioned on A according to the random sampling model calculated by the method described in Section 2.1, and the blue lines show the 4-fold local SD of M estimated by the comparison of technical replicates (as described in Section 2.2). See Supplementary Methods Section 3 for detail. (**C**) An example of differentially expressed genes (red points) identified between kidney and liver by the MA-plot-based method with random sampling model at an FDR of 0.1%. The red lines show the 'theoretical' 4-fold local SD of M according to the random sampling model. (**D**) An example of differentially expressed genes (red points) identified between kidney and liver by MA-plot-based method with technical replicates at an FDR of 0.1%. Blue points are from the replicates (kidneyR1L1 and kidneyR1L3), and the blue lines show the 4-fold local SD of M for the two technical replicates.

to identify the difference of the gene expression between the two samples (see Supplementary 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 and Hochberg, 1995; Storey and Tibshirani, 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 versus 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 SD of repeated measurements and uses permutations of the repeated measurements to estimate FDR.

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 3 pM) were tested for each sample. We used the refFlat gene annotation file downloaded from UCSC Genome browser and chose the method proposed by Storey and Tibshirani (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 (Supplementary 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 dataset have little technical variation (Marioni *et al.*, 2008). And none of the gene was falsely identified as differentially expressed between the two replicates by each method at an FDR of 0.1%, respectively (Supplementary Table 1). However, samples sequenced at different concentrations showed larger variance (Supplementary 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 an FDR of 0.1%, respectively. And the lists of differentially expressed genes given by different methods are quite consistent with each other (Supplementary Table S2). Figure 1C and 1D shows the results given by the MA-plot-based method with random sampling model and with technical replicates, respectively. And Supplementary 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 binominal distribution to model overdispersion relative to Poisson for digital gene expression data with small replicates (Robinson and Smyth, 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.

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

Conflict of Interest: none declared.

REFERENCES

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*, **57**, 289–300.

Bloom,J.S. *et al.* (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, **10**, 221.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

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

Robinson,M,D. (2009) edgeR: empirical analysis of digital gene expression data in R. Available at http://bioconductor.org/packages/2.4/bioc/html/edgeR.html (last accessed date September 16, 2009).

Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Storey,J. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Tibshirani,R. *et al.* (2009) samr: SAM: significance analysis of microarrays. Available at http://cran.r-project.org/web/packages/samr/index.html (last accessed date September 16, 2009).

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Supplemental Material

This Supplemental Material provides additional details for the methods presented in DEGseq and supplemental results given by the methods that we performed on the data from Marioni *et al.* (2008).

Contents

Supplemental Methods

1. MA-plot-based method with random sampling model

1.1 The conditional distribution of *M* given *A*

Let C_1 and C_2 denote the counts of reads mapped to a specific gene obtained from two samples with $C_i \sim$ binomial(n_i, p_i), $i = 1, 2$, where n_i denotes the total number of mapped reads and p_i denotes the probability of a read coming from that gene. We define $M = \log_2 C_1 - \log_2 C_2$, and $A = (\log_2 C_1 + \log_2 C_2)/2$. We assume that C_1 and C_2 are independent. Let $X = \log_2 C_1$ and $Y = \log_2 C_2$, hence $M = X - Y$ and $A = (X + Y)/2$.

When n_1 and n_2 are large enough, we can obtain the asymptotic distribution of C_1/n_1 and C_2/n_2 as follows:

$$\sqrt{n_1}(\frac{C_1}{n_1} - p_1) \to N(0, p_1(1-p_1)) ,$$

$$\sqrt{n_2}(\frac{C_2}{n_2} - p_2) \to N(0, p_2(1-p_2)) .$$

Then $X = g(C_1/n_1) = \log_2(n_1 C_1/n_1)$, where $g(x) = \log_2(n_1 x)$. According to Delta Method, we can obtain the asymptotic distribution of X when $n_1 \to \infty$, that is

$$\begin{aligned}\sqrt{n_1}(X - \log_2(n_1 p_1)) &= \sqrt{n_1}(g(\frac{C_1}{n_1}) - g(p_1)) \\ &\to N(0, p_1(1-p_1)[g'(p_1)]^2) \\ &= N(0, (\frac{1-p_1}{p_1})(\log_2 e)^2) .\end{aligned}$$

Thus,

$$X \to N(\log_2(n_1 p_1), (\frac{1-p_1}{n_1 p_1})(\log_2 e)^2) .$$

Similarly, the asymptotic distribution of Y can be obtained:

$$Y \to N(\log_2(n_2 p_2), (\frac{1-p_2}{n_2 p_2})(\log_2 e)^2) .$$

Thus, we have proved X and Y follow normal distributions approximately (when n_i is large enough), and denote

$$X \sim N(\mu_X, \sigma_X^2) ,$$
$$Y \sim N(\mu_Y, \sigma_Y^2) .$$

Based on the assumption that C_1 and C_2 are independent (so X and Y are independent), the distributions of M and A can be obtained:

$$M \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) = N(\mu_M, \sigma_M^2) ,$$

$$A \sim N(\frac{1}{2}(\mu_X + \mu_Y), \frac{1}{4}(\sigma_X^2 + \sigma_Y^2)) = N(\mu_A, \sigma_A^2) .$$

Then, the conditional distribution of M given that $A=a$ can be obtained:

$$M\,|(A=a) \sim N(\mu_M + \rho\frac{\sigma_M}{\sigma_A}(a-\mu_A), \sigma_M^2(1-\rho^2)),$$

where ρ is the correlation coefficient between M and A.

The covariance between M and A is

$$Cov(M,A) = E(MA) - \mu_M\mu_A = \frac{1}{2}E(X^2 - Y^2) - \frac{1}{2}(\mu_X^2 - \mu_Y^2) = \frac{1}{2}(\sigma_X^2 - \sigma_Y^2),$$

so $$\rho = \frac{Cov(M,A)}{\sigma_M\sigma_A} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}.$$

Thus,

$$E(M\,|A=a) = \mu_M + \rho\frac{\sigma_M}{\sigma_A}(a-\mu_A)$$

$$= \mu_X - \mu_Y + 2\frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}(a - \frac{1}{2}(\mu_X + \mu_Y)),$$

and

$$Var(M\,|A=a) = \sigma_M^2(1-\rho^2) = \frac{4\sigma_X^2\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}.$$

1.2 Hypothesis test based on the random sampling model

For gene g with (A=a, M=m) on the MA-plot of two samples, we do the hypothesis test H_0: $p_1 = p_2 = p$ versus H_1: $p_1 \neq p_2$.

Based on above deduction,

$$\mu_A = \frac{1}{2}(\mu_X + \mu_Y) = \frac{1}{2}\log_2(n_1 n_2 p^2).$$

Thus,

$$p = \sqrt{2^{2\mu_A}/(n_1 n_2)}.$$

Use a as an estimate of μ_A then

$$\hat{p} = \sqrt{2^{2a}/(n_1 n_2)}.$$

So the estimates of $E(M \mid A$=a) and $Var(M \mid A$=a) are

$$\hat{E}(M \mid A=a) = \log_2(n_1) - \log_2(n_2),$$

and

$$\hat{Var}(M \mid A=a) = \frac{4(1-\sqrt{2^{2a}/(n_1 n_2)})(\log_2(e))^2}{(n_1+n_2)\sqrt{2^{2a}/(n_1 n_2)}}.$$

Then use the two estimates to calculate a Z-score for the gene g with (A=a, M=m), and convert it to a two-sided P-value which is used to indicate whether gene g is differentially expressed or not.

$$Z - score = \frac{|\,m - \hat{E}(M \mid A=a)\,|}{(\hat{Var}(M \mid A=a))^{\frac{1}{2}}}.$$

Given a Z-score threshold, take four as an example, the two lines with the following equations are used to indicate the four-fold local standard deviation of M according to the random sampling model:

$$m_1 = \hat{E}(M \mid A = a) + 4 * (\hat{V}ar(M \mid A = a))^{\frac{1}{2}}$$
$$= \log_2(n_1) - \log_2(n_2) + 4 * (\frac{4(1 - \sqrt{2^{2a} / (n_1 n_2)})(\log_2(e))^2}{(n_1 + n_2)\sqrt{2^{2a} / (n_1 n_2)}})^{\frac{1}{2}}$$

$$m_2 = \hat{E}(M \mid A = a) - 4 * (\hat{V}ar(M \mid A = a))^{\frac{1}{2}}$$
$$= \log_2(n_1) - \log_2(n_2) - 4 * (\frac{4(1 - \sqrt{2^{2a} / (n_1 n_2)})(\log_2(e))^2}{(n_1 + n_2)\sqrt{2^{2a} / (n_1 n_2)}})^{\frac{1}{2}}$$

We call the lines obtained by the above equations "theoretical" four-fold local standard deviations lines. See the red lines in Figure 1B and Figure 1C as examples.

2. MA-plot-based method with technical replicates

To estimate the noise level of genes with different intensity, and identify gene expression difference in different sequencing libraries, we employed this statistical model based on the MA-plot. Here M is the Y-axis and represents the intensity ratio, and A is the X-axis and represents the average intensity for each transcript. To estimate the random variation, we first draw a MA-plot using two technical replicates (e.g. two sequencing lanes) from the same library. A sliding window is applied to scan the MA-plot along the A-axis (see Supplemental Fig. S1D). In fact, the windows are much narrower than those in the Figure S1D. Each window includes 1% points of the MA-plot. To get the local standard deviation of each window i (i=1,2,…,100), we calculate the mean μ_i and standard deviation σ_i of M of all the points in the window. Suppose the window i is centered at A=a_i. Given a Z-score threshold, take four as an example, we can draw two points (a_i, μ_i+4*σ_i) and (a_i, μ_i-4*σ_i) for each window i (the blue points in Supplemental Fig. S1D). Next, the two set of points (a_i, μ_i+4*σ_i) and (a_i, μ_i-4*σ_i) are regressed to two blue lines. The blue lines are then used to predict the local mean μ_a and standard deviation σ_a of M for A=a for the two technical replicates. We call the two blue lines the four-fold local standard deviation lines of the two technical replicates.

To identify differentially expressed genes between two different samples, we draw a second MA-plot for the data from the two samples. For each transcript g with (A=a_g, M=m_g) on the MA-plot, we use the two blue lines (got by above steps) predict the local mean μ_g and standard deviation σ_g of M for A=a_g under the null hypothesis: gene g has the same expression value between the two samples. Then a Z-score=$|m_g$-$\mu_g|/\sigma_g$ is calculated under the assumption of normal distribution. Finally, a P-value is assigned to this gene according to the Z-score to evaluate whether this gene is differentially expressed.

3. Method to check whether the variation between technical replicates can be explained by the random sampling model

To check the variation between technical replicates, we draw the "theoretical" four-fold local standard deviation lines (red lines in Fig 1B and Supplemental Fig. S1A) and the four-fold local standard deviation lines estimated by the comparison of technical replicates (blue lines in Fig 1B and Supplemental Fig. S1A). If the lines are coincidence with each other, we can say that the variation between technical replicates can be fully explained by the random sampling model. Otherwise, there is background noise for the data of the two technical replicates. For example, on the plot of technical replicates kidneyR1L1 and kidneyR1L3 (see Fig. 1B), the red lines and the blue lines are almost coincidence with each other. This indicates the two replicates have little technical variation. While for the replicates kidneyR1L1 and kidneyR2L4 which are sequenced at different concentrations (see Supplemental Fig. S1A), it can be inferred that the two replicates could not be fully explained by the random sampling model.

Supplemental Figures

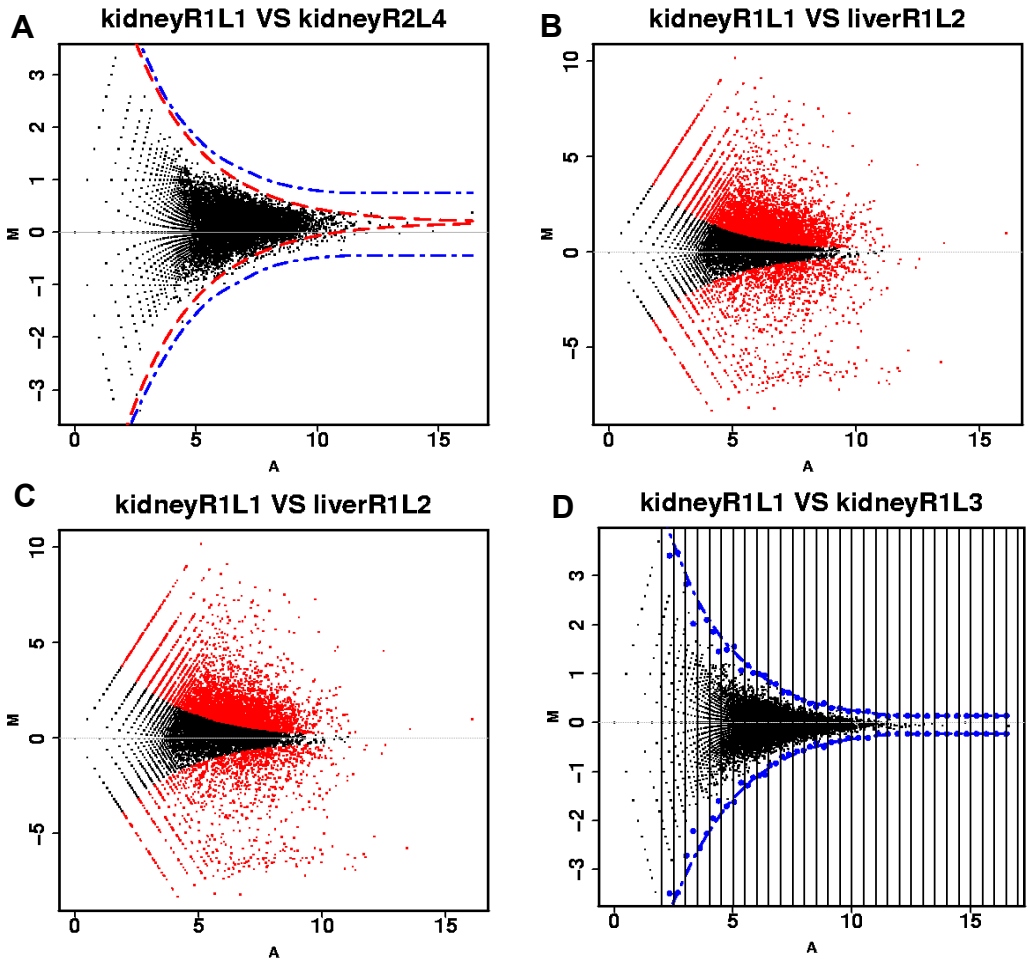

Fig. S1. (**A**) The plot generated by DEGseq showing whether the variation between technical replicates can be largely explained by the random sampling model. The two red lines correspond to the "theoretical" four-fold local standard deviation of *M* according to the random sampling model, and the blue lines show the four fold local deviation of *M* estimated by the comparison of technical replicates. The technical replicates kidneyR1L1 and kidneyR2L4 were generated at different cDNA concentrations. (**B**) An example of differentially expressed genes (red points) identified between kidney and liver using the likelihood ratio test that was used by Marioni *et al.* (2008) at a FDR of 0.1%. (**C**) An example of differentially expressed genes (red points) identified between kidney and liver using the Fisher's exact test that was used by Bloom *et al*. (2009) at a FDR of 0.1%. (D) The example MA-plot for explaining how to get the local standard deviation between two replicates.

Supplemental Tables

Table S1. The number of genes identified by the four methods between kidneyR1L1 and kidneyR1L3 and between kidneyR1L1 and liverR1L2 at FDR 0.1%.

Method	kidneyR1L1 and kidneyR1L3	kidneyR1L1 and liverR1L2
FET	0	6357
LRT	0	6966
MARS	0	6857
MATR	-	6648

We use FET, LRT, MARS and MATR stand for the method using Fisher's exact test, the method using likelihood ratio test, the MA-plot-based method with random sampling model, and the MA-plot-based method with technical replicates. The method MATR used technical replicates kidneyR1L1 and kidneyR1L3 when identifying differentially expressed genes between kidneyR1L1 and liverR1L2. All the methods took the total count of reads that uniquely mapped to genome as the depths of the samples. The gene expression values are calculated with the gene annotation file refFlat.txt downloaded from http://genome.ucsc.edu.

Table S2. The number of identical genes identified by each two methods between kidneyR1L1 and liverR1L2 at FDR 0.1%.

Method	FET	LRT	MARS	MATR
FET	-	6357	6357	6357
LRT	6357	-	6857	6648
MARS	6357	6857	-	6636
MATR	6357	6648	6636	-

We use FET, LRT, MARS and MATR stand for the method using Fisher's exact test, the method using likelihood ratio test, the MA-plot-based method with random sampling model, and the MA-plot-based method with technical replicates. The method MATR used technical replicates kidneyR1L1 and kidneyR1L3 when identifying differentially expressed genes between kidneyR1L1 and liverR1L2. All the methods took the total count of reads that uniquely mapped to genome as the depths of the samples. The gene expression values are calculated with the gene annotation file refFlat.txt downloaded from http://genome.ucsc.edu.

REFERENCES

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-seq. *Bioinformatics*, **25**, 1026–1032.

Bloom,J.S. *et al.* (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, **10**, 221.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Storey,J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440-9445.