Cell & Bioscience

REVIEW Open Access

Statistical methods for identifying differentially expressed genes in RNA-Seq experiments

Zhide Fang[1*], Jeffrey Martin[2,3] and Zhong Wang[2,3,4]

Abstract

RNA sequencing (RNA-Seq) is rapidly replacing microarrays for profiling gene expression with much improved accuracy and sensitivity. One of the most common questions in a typical gene profiling experiment is how to identify a set of transcripts that are differentially expressed between different experimental conditions. Some of the statistical methods developed for microarray data analysis can be applied to RNA-Seq data with or without modifications. Recently several additional methods have been developed specifically for RNA-Seq data sets. This review attempts to give an in-depth review of these statistical methods, with the goal of providing a comprehensive guide when choosing appropriate metrics for RNA-Seq statistical analyses.

Introduction

Transcriptomics holds the key to understanding how the information encoded in the genome is translated into cellular functions, and how this translation process responds to the changing environment. Given a transcriptome, or the collection of all the transcripts including both protein coding mRNAs and noncoding RNAs, one of the outstanding questions in transcriptomics is to accurately quantify the abundance of each transcript within different tissues and time points, and to correlate changes in abundance to genetic and environmental perturbation in order to understand genome function and adaptation.

Transcriptome profiling, or gene expression profiling, is the technology used to determine the steady state abundance of each transcript within a transcriptome. Transcriptome profiling is traditionally done using either quantitative RT-PCR (reviewed in [1]) to interrogate a few genes, or microarrays (cDNA array [2] or whole genome tiling array [3-5]) to investigate genome-wide transcriptional activity. Recently, as a result of the low cost of next generation sequencing technologies [6], transcriptome profiling by RNA sequencing (RNA-Seq) is becoming the method of choice because of its unprecedented sensitivity and accuracy (reviewed in [7,8]). Unlike prior technologies, next-generation sequencing technologies allow reference transcriptomes to be assembled directly from RNA-Seq data, thereby eliminating the need for existing reference genomes or transcriptomes [9]. This capability is particularly attractive for non-model organisms or microbial communities that lack high quality references.

There are both shared and unique aspects in the experimental design and data generation phases of expression microarrays and RNA-Seq technologies, and these attributes are compared elsewhere [7,8,10,11]. For data analysis there are three major steps for both technologies: data preprocessing, statistical analysis and functional interpretation (Figure 1). Preprocessing microarray data normally includes background correction, normalization and summation, while preprocessing RNA-Seq data includes artifact filtering and short read alignment/assembly. The bioinformatics details involved in preprocessing microarray and RNA-Seq data have been reviewed previously [8,12,13]. After data preprocessing the expression level of each transcript is determined. For microarrays the levels are often represented as continuous numbers, while for RNA-Seq datasets expression levels are represented as discrete read counts. Statistical analysis is then performed to identify differentially expressed transcripts among different samples/conditions, and the results can be further analyzed to gain functional insights (Figure 1). In this review we focus on the statistical methods that are used to identify differentially expressed transcripts in RNA-Seq experiments. Some of them (for

* Correspondence: zfang@lsuhsc.edu
[1]Biostatistics Program, School of Public Health, LSU Health Sciences Center, 2020 Gravier Street, 3rd floor, New Orleans, LA 70112, USA
Full list of author information is available at the end of the article

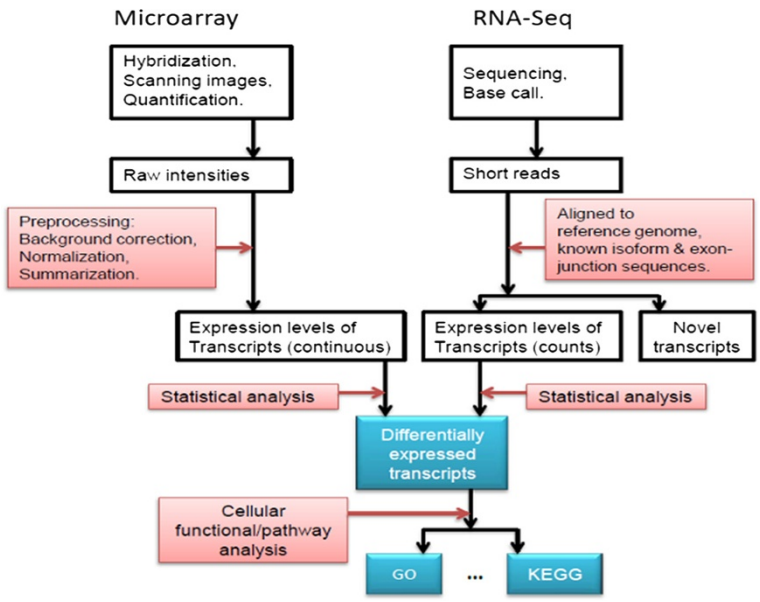

Figure 1 The analysis workflows of microarray and RNA-Seq data.

example, likelihood ratio test) have been used for microarray data analysis and then were adapted for RNA-Seq data analysis, while others were developed specifically for RNA-Seq.

Over the past decade, various statistical analysis tools have been developed to analyze expression profiling data generated by microarrays (Reviewed in [14]). Before these tools can be applied to RNA-Seq data, it is worth noting that microarray data and RNA-Seq data are inherently different. As mentioned earlier, microarray data is "analog" since expression levels are represented as continuous hybridization signal intensities. In contrast, RNA-Seq data is "digital", representing expression levels as discrete counts. This inherent difference leads to the difference in the parametric statistical methods that are used since they often depend on the assumptions of the random mechanism that generates the data. For example, the normal distribution is a common distribution for statistical comparisons involving continuous data. It is generally assumed that the log intensities (or expression levels) in a microarray experiment are approximately normally distributed. In contrast, the Poisson, Binomial and Multinomial distributions are more suitable for modeling discrete data in an RNA-Seq experiment. Therefore a statistical method developed for microarray data analysis cannot be directly applied to RNA-Seq data analysis without first examining the underlying distributions. Recently several statistical methods have been developed to deal specifically with RNA-Seq count data [14-17]. In this review we summarize these methods, while focusing on the pros and cons of each method in the context of specific applications.

RNA-Seq count data

As mentioned previously, the expression levels measured by RNA-Seq experiments are represented by the number of reads derived from each transcript in a transcriptome. We will not discuss the problem of resolving the expression levels of alternatively spliced transcript isoforms from a single gene, since this is still a challenge and undergoing active research [7,18-20]. Here, for simplicity, we use the term "gene" or "transcript" to generically refer to a spliced mRNA isoform, a non-coding RNA, a small RNA, or any product resulted from a transcriptional, splicing, or post-transcription-modification event.

RNA-Seq count data can be organized into a numerical $(p \times n)$ matrix (M), with p representing the number of genes and n the number of samples. We use phenotype to refer to an experimental condition, treatment, tissue, or time point. Typically the number of genes is far greater than the number of samples $(p \gg n)$. Another dimension often present in RNA-Seq datasets is the number of replicates. Since the RNA-Seq protocol is highly reproducible [21-23], technical replicates are usually not necessary, and instead 2–3 biological replicates are used to reduce the degree of noise resulting from biological variations. Generally we assume that there are n_k biological (or technical) replicates under the k^{th} phenotype, $k = 1, 2, \cdots, t$. As a result, a typical RNA-Seq

data set has a series of two-dimensional sub-matrices containing non-negative integers (counts), with each sub-matrix being derived from a specific phenotype. Therefore, we have $n = \sum_{k=1}^{t} n_k$, and the data matrix can be arranged As M below, with the element $m_{ij}^{(k)}$ being the expression level of the i^{th} gene from the j^{th} replicate in the kth phenotype, $i = 1, 2, \cdots, p, j = 1, 2, \cdots, n_k$, $k = 1, 2, \cdots, t$.

RNA-Seq count data normalization

An important consideration to make prior to statistical analysis is normalization. The sequencing depth, or library size, which is usually defined as the total number of aligned sequences in each sample, often varies from one sample to another. Denote $L_j^{(k)}$ as the sequencing depth for the j^{th} sample in the k^{th} phenotype. Then we have $L_j^{(k)} = \sum_{i \geq 1} m_{ij}^{(k)}$. Normalizing count data transforms it from discrete to continuous. For example, the RPKM metric (Reads Per Kilobase of transcript per Million mapped reads [22]) is used to measure the relative expression level of a transcript. Although RPKM considers the length of the transcript, and thus allows for comparison among different transcripts, in most studies the gene length is not an issue because the comparison is made for the same gene between different conditions [24]. RPKM-based expression measurements cannot be directly used for the count-based models.

In this review we assume that the data in the matrix M are raw read counts without normalization. Unlike microarray data analysis which often requires sophisticated normalization procedures to compensate for biases introduced from sample loading, imaging, and other technical or biological factors [25], RNA-Seq data typically does not require a separate normalization step for two reasons: 1) The difference in the sequencing depths or library sizes between different samples is addressed through the parameterization of the underlying distributions (see below). 2) Some models already take into account the variation among biological replicates. For example, the over-dispersion parameter (discussed below) in the Negative Binomial model accounts for the variation across the biological replicates [26].

In the next section we will discuss the statistical methods that have been developed to address whether or not a gene is differentially expressed among a group of time points or conditions. In the case of t = 2, this problem reduces to the two-group (pair-wise) comparison.

Statistical methods to detect differentially expressed genes

Several statistical methods have been proposed to detect the differentially expressed genes from a counts table (Table 1). The number of samples or replicates in a typical RNA-Seq experiment is usually small, thereby excluding the application of nonparametric methods that implement sample permutations. For this reason, here we focus on parametric methods only. These methods differ in their underlying data distributions, how they handle biological replicates, and their ability to perform multi-group comparisons. Some of these methods have been implemented in related *R/Bioconductor* packages (Table 1). Each of these methods is discussed in further detail below.

Methods based on the poisson distribution

In an RNA-Seq dataset, the expression level of a specific gene, $m_{ij}^{(k)}$, is defined as the total number of short sequences which aligned to the gene. That is, it is the sum of a series of random events. Each event corresponds to a short sequence and follows a Bernoulli distribution with the probability of success equating to the probability that the sequence aligns to the gene. Since the read alignments can be assumed to be independent, the distribution of $m_{ij}^{(k)}$ can be approximated by a Poisson distribution, $Poi(\mu_i^{(k)})$, with $\mu_i^{(k)}$ being the mean. This Poisson model is verified in the case where there are only technical replicates using a single source of RNA [21]. For the ith gene, the statistical null hypothesis in testing different expression levels across phenotypes is that all of the means are equal. The statistical test procedures based upon Poisson modeling are reviewed in the next subsection.

Fisher's exact test

This method can be used for comparing two phenotypes (t = 2). For the ith gene, we can form a 2x2 contingency table for its expression values in the matrix M [27], Table 2.

The Fisher's exact test is to test whether or not there exists a significant association between the gene expression and the phenotype, in other words, whether or not

$$M = \left(\begin{bmatrix} m_{11}^{(1)} & m_{12}^{(1)} & \cdots & m_{1\,n_1}^{(1)} \\ m_{21}^{(1)} & m_{22}^{(1)} & \cdots & m_{2\,n_1}^{(1)} \\ & \vdots & & \\ m_{p1}^{(1)} & m_{p2}^{(1)} & \cdots & m_{p\,n_1}^{(1)} \end{bmatrix} \begin{bmatrix} m_{11}^{(2)} & m_{12}^{(2)} & \cdots & m_{1\,n_2}^{(2)} \\ m_{21}^{(2)} & m_{22}^{(2)} & \cdots & m_{2\,n_2}^{(2)} \\ & \vdots & & \\ m_{p1}^{(2)} & m_{p2}^{(2)} & \cdots & m_{p\,n_2}^{(2)} \end{bmatrix} \cdots \begin{bmatrix} m_{11}^{(t)} & m_{12}^{(t)} & \cdots & m_{1\,n_t}^{(t)} \\ m_{21}^{(t)} & m_{22}^{(t)} & \cdots & m_{2\,n_t}^{(t)} \\ & \vdots & & \\ m_{p1}^{(t)} & m_{p2}^{(t)} & \cdots & m_{p\,n_t}^{(t)} \end{bmatrix} \right)$$

Table 1 A comparison of common statistical methods for RNA-Seq differential gene expression analysis

Method	Underlying distribution	Recommended with biological replicates	Multi-group comparison	R/Bioconductor package	Reference
Fisher's exact Test	Poisson	No	No	No	[27]
Likelihood ratio test	Poisson	No	Yes	No	[21]
edgeR	Negative Binomial	Yes	Yes	Yes	[28]
DESeq	Negative Binomial	Yes	No	Yes	[15]
baySeq	Negative Binomial	Yes	Yes	Yes	[17]
BBSeq	Beta-Binomial	Yes	No	Yes	[29]
Two-stage poisson model	Poisson	Yes	Yes	No	[30]

the odds ratio is significantly greater or less than 1. This test is based on the fact that with the assumption of Poisson sampling and fixed marginal totals, the count $m_{i1}^{(1)}$ follows a hyper-geometric distribution. The p value is the total of the hyper-geometric probabilities for outcomes at least as favorable to the alternative hypothesis (the gene expression in phenotype 1 is lower than that in phenotype 2 (the odds ratio is < 1), or vice versa) as the observed outcome [31]. A simple R function,

$$> x = matrix\left(cm\binom{(1)}{i1}, m_{i1}^{(2)}, L_1^{(1)} - m_{i1}^{(1)}, L_1^{(2)} - m_{i1}^{(2)}\right), nrow = 2)$$

$$> fisher.test(x, alternative = c\left("two.sided", "greater", "less"\right))$$

will give the p value of the test.

Since the null hypothesis of independence in Fisher's exact test is equivalent to the null hypothesis that the odds ratio is equal to 1, one can avoid a potential false positive due to the difference in the sequencing depths. As an example, consider the case: $m_{i1}^{(1)} = 10, m_{i1}^{(2)} = 20$, $L_1^{(1)} = 1e+6, L_1^{(2)} = 2e+6$. The estimated odds ratio is 1 (no association by Fisher's exact test), but the fold change is 2 (possibly declared as differential by other tests based on fold change). Furthermore, though Fisher's exact test was designed for analyzing datasets without replicates, if there are replicates and Poisson sampling holds true, the test can still be applied – one simply sums up the replicates under the same condition to form the 2×2 contingency table.

The p value obtained above is for a single gene. As in the analysis of expression data from microarray experiments, there are thousands of genes in one RNA-Seq experiment and thus we need to consider the problem of an inflated false positive rate due to multiple hypothesis testing. This problem can be addressed by directly adjusting p values or calculating q values [32,33]. Many methods have been proposed to calculate adjusted p-values, including Bonferroni's single-step adjusted p-values, Holm's step-down adjusted p values [34], Hochberg's step-up adjusted p-values [35], and many more. The R function mt.raw2adjp() in the R/Bioconductor package multtest computes adjusted p values, with nine different computing procedures. Q values can be obtained by the R function qvalue() in the R/Bioconductor package qvalue. All of these functions take the vector of raw p values as the input argument.

Table 2 The 2×2 contingency table for one gene

	Gene	Not Gene
Phenotype 1	$m_{il}^{(1)}$	$L_l^{(1)} - m_{il}^{(1)}$
Phenotype 2	$m_{il}^{(2)}$	$L_l^{(2)} - m_{il}^{(2)}$

Likelihood ratio test

Marioni et al. assumed that the gene count $m_{ij}^{(k)}$ follows a Poisson distribution, $Poi(\mu_i^{(k)} = L_j^{(k)} v_i^{(k)})$, where $v_i^{(k)}$ represents the proportion of gene transcript copies of the ith gene in all samples under the kth phenotype (k = 1, 2, for pair-wise comparison), and then used the likelihood ratio test to identify the differentially expressed genes [21]. The purpose of incorporating the sequencing depth ($L_j^{(k)}$) parameter into the Poisson mean is to reduce the variation in sequencing depth. If we look into the significance of differential expressions of genes on a gene-by-gene basis, the likelihood function and maximum likelihood estimations are easy to obtain. For the two-sided alternative hypothesis, by applying simple algebraic operations we have the likelihood ratio test statistic,

$$\Lambda_i = -2\left(\left(\sum_j m_{ij}^{(1)}\right) \times \log\left(\frac{\sum_j m_{ij}^{(1)} + \sum_j m_{ij}^{(2)}}{\sum_j m_{ij}^{(1)}} \frac{\sum_j L_j^{(1)}}{\sum_j L_j^{(1)} + \sum_j L_j^{(2)}}\right) + \left(\sum_j m_{ij}^{(2)}\right) \times \log\left(\frac{\sum_j m_{ij}^{(1)} + \sum_j m_{ij}^{(2)}}{\sum_j m_{ij}^{(2)}} \frac{\sum_j L_j^{(2)}}{\sum_j L_j^{(1)} + \sum_j L_j^{(2)}}\right)\right),$$

And the p value for an individual gene can be calculated as the right tailed probability of a Chi-squared distribution with 1 degree of freedom. For the one-sided alternative hypothesis, $v_i^{(1)} > v_i^{(2)}$ the p value is half of the above right-tailed probability of the Chi-squared distribution if the unconstrained maximum likelihood estimates of $v_i^{(1)}, v_i^{(2)}$ satisfy: $\left(\hat{v}_i^{(1)} = \right) \frac{\sum_j m_{ij}^{(1)}}{\sum_j L_j^{(1)}} > \frac{\sum_j m_{ij}^{(2)}}{\sum_j L_j^{(2)}} \left(= \hat{v}_i^{(2)}\right)$; or 0.5 otherwise [36]. Adjusted p values or q values to control the false positive rate can be obtained by the methods described in the previous subsection.

Poisson modeling is an appropriate fit not only for sequencing data with technical replicates [21], but also for those with biological replicates, as long as the sample mean is close to the sample variance. However, the requirement that the variance is the same as the mean excludes the application of the Poisson model to RNA-Seq data, should over-dispersion (defined below) occur. The likelihood ratio test may give misleading results if the assumptions about the sampling distribution are violated.

Models for over-dispersed count data

Given a sampling distribution, over-dispersion occurs if the observed variance is greater than the assumed variance. In the Poisson model, over-dispersion occurs if the sample variance is greater than the sample mean. There could be several sources that cause over-dispersion in RNA-Seq data, including the variability in biological replicates due to heterogeneity within a population of cells, possible correlation between gene expressions due to regulation, and other uncontrolled variations. The existence of over-dispersion in real data was observed in several previous studies [26,30]. Popular models to safeguard against over-dispersion include the negative binomial distribution, beta-binomial distribution or two-stage Poisson distribution, as discussed below.

Negative binomial model

As mentioned above, when over-dispersion is observed across the samples, the gene counts cannot be estimated accurately by a simple Poisson model. One way to handle this problem is to apply a Bayesian method – allowing the Poisson mean to be a random variable and then model the gene counts by the marginal distribution of $m_{ij}^{(k)}$. Specifically, assume that the Poisson mean follows a Gamma distribution with the scale parameter $\mu_i^{(k)}\phi$ and the shape parameter $\left(1/\phi\right)$, then the marginal distribution of the gene count has a Negative Binomial distribution with mean $\mu_i^{(k)}$ and the variance $\mu_i^{(k)}(1 + \mu_i^{(k)}\phi)$ [37]. The Negative Binomial distribution can model the over-dispersed Poisson gene count where $\phi > 0$ and reduces to the Poisson distribution as $\phi \to 0$. The R/Bioconductor package "edgeR" applies this model to detect the differentially expressed genes in RNA-Seq data [28], where the mean of the Negative Binomial is rewritten as $(\mu_i^{(k)} = L_j^{(k)} v_i^{(k)})$ to adjust for the difference in sequencing depths across the samples. The ith gene is differentially expressed if the parameters $v_i^{(k)}$ are significantly different across phenotypes. For simple, pair-wise comparisons between phenotypes, the Negative Binomial parameters are estimated by conditional maximum likelihood and quantile-adjusted conditional maximum likelihood [26,37], and then an exact test (similar to Fisher's exact test) is carried out to generate p values for individual genes. These can be completed by using the R function "exactTest()" with options for different estimates of the dispersion. For complex, mutligroup comparisons among phenotypes, edgeR applies the Cox-Reid profile-adjusted likelihood method to estimate the Negative Binomial parameters [38], and then uses the generalized linear model likelihood ratio test (R functions glmFit() and glmLRT()) to discover differentially expressed genes.

While the relationship between the Negative Binomial mean and variance simplifies the estimation of these parameters, it may result in some variation unexplained by the model and thus potentially introduce selection biases in the differentially expressed genes. Anders and Huber extended the method in edgeR by hierarchically modeling this mean-variance relationship [15]. Their method is implemented in a R/Bioconductor package, called DESeq. The R function nbinomTest() or nbinomTestForMatrics() can return unadjusted p values for individual genes. The adjusted p values from the "BH" method [39] are also given by the first function. To our knowledge, DESeq is currently limited to pair-wise comparisons.

Another modification to edgeR is given by Hardcastle and Kelly [17]. Their method is based on an empirically Bayesian approach and can be used for multi-group comparisons as well as for pair-wise comparisons. The gene count is also modeled by the Negative Binomial distribution. For each gene, instead of calculating a p value, a posterior probability is obtained for each comparison (alternative) among phenotypes. The probability of differential expression of a gene is defined as the sum of the posterior probabilities for all possible comparisons. Then, the genes are ranked based upon the probability of differential expression This method is implemented in the R/Bioconductor package, baySeq. The R function getPosteriors() returns the posterior likelihoods of comparisons. One of the disadvantages of this method is that it is more computationally expensive since all types of comparisons (alternatives) are considered and the number of comparisons increases dramatically when the number of phenotypes increases.

Beta-binomial model

Zhou, Xia and Wright model the gene count in a sample with a Beta-Binomial distribution [29]. Assuming that whether or not a short sequence is mapped to a particular transcript follows a Bernoulli law with a mapping probability θ, then θ has a prior of the Beta distribution. The introduction of randomness to the mapping probability is to account for the over-dispersion in the gene count data, with the over-dispersion being explained by a Beta distribution parameter ϕ. The maximum likelihood estimates of parameters (E(θ), ϕ) are obtained either by a free model in which both parameters are unrelated or by a constrained model in which a mean-overdispersion relationship is assumed. A logistic model is fitted to model the relationship between the mean E(θ) and the design matrix of covariates including the indicator variables for phenotypes. Then the significance of an indicator variable is determined by a Wald statistic (the ratio of the corresponding coefficient in the fitted logistic model and its standard error) and indicates whether or not the gene is differentially expressed. This method is implemented in the R package, BBSeq, which is available on this website (last accessed on 03/07/2012): http://www.bios.unc.edu/research/genomic_software/BBSeq/. Two R functions, *free.estimate() and constrained.estimate()*, can generate the raw p values for genes in pair-wise comparisons. However, no function in the package directly gives the p values for multi-group comparisons.

The Beta-Binomial model has been widely used to model the over-dispersed, discrete count data. For example, it was applied to analyze tag count data derived from the Serial Analysis of Gene Expression (SAGE) [40]; tag count data obtained from Digital Gene Expression profiling [41]; and spectral count data generated from label-free tandem mass spectrometry-based proteomic experiments [42]. To model RNA-Seq data with a Beta-Binomial distribution, the probability that a short sequence is mapped to a specific transcript is implicitly assumed to be constant for all short sequences in a sample. We have not seen any verification or justification of this assumption in the literature.

Two-stage poisson model

Auer and Doerge [30] proposed a Two-Stage Poisson Model for RNA-Seq data analysis, based upon the argument that the over-dispersion is most likely caused by within group variation in expression if the experiment includes independent biological replicates without a significant population structure. The method consists of two steps. In the first step, genes are classified into two exclusive classes, genes with or without over-dispersion, by using an adjusted score test on the hypothesis of whether or not the over-dispersion is present within the count data of a gene. Then in the second step, for genes without significant over-dispersion, differential expression is tested by a standard likelihood approach with maximum likelihood estimates being obtained under the Poisson model. Raw p values are calculated by an approximated Chi-squared distribution of degree one. For genes with significant over-dispersion, they use the quasi-likelihood statistic that is defined as the ratio of the likelihood statistic and an estimate of over-dispersion. Raw p values are calculated from an *F*-distribution. The built-in R functions "*glm()*" and "*deviance()*" can be used to obtain the likelihood ratio statistics. The detailed R code for p values can be downloaded from the author's website (http://www.stat.purdue.edu/~doerge/software/TSPM.R), last accessed on 03/07/2012.

Under the model assumptions, the authors demonstrated that the Two-Stage Poisson Model is a powerful

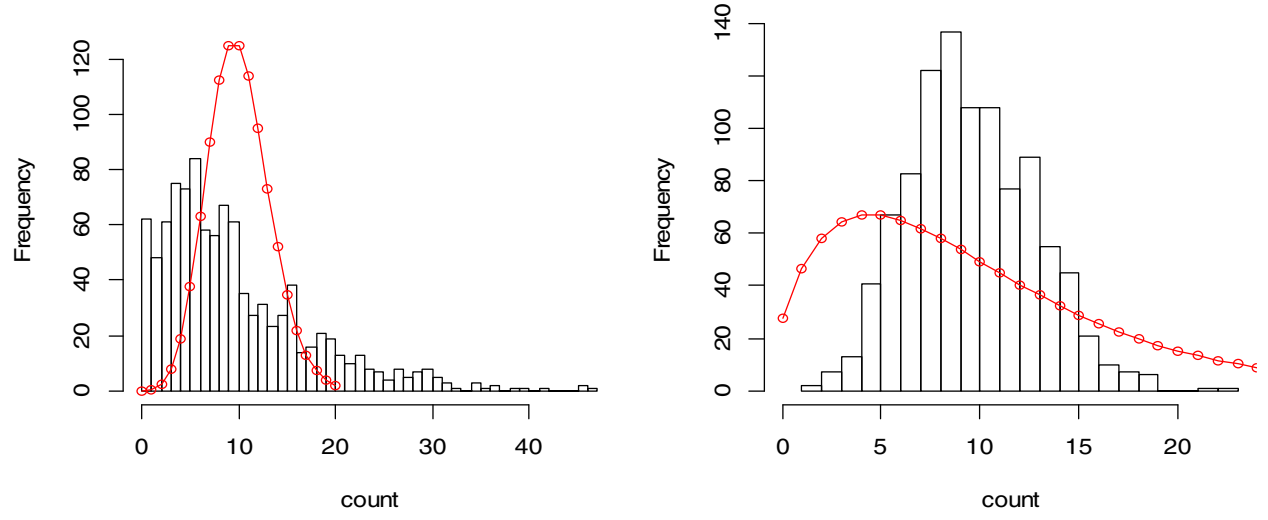

Figure 2 Histograms and wrongly fitted models for 1000 simulated data points.

tool for detecting differentially expressed genes. However, if other confounding factors exist such that the levels of transcription within a phenotype are not stable, this method may not control the false positive rate well. Furthermore, as pointed out by the authors, this method requires a relatively higher degree of freedom (the difference between the sample size and the number of phenotypes) in order to be effective.

Conclusion and future perspectives

Next-generation sequencing technologies are revolutionizing genomic/proteomic studies, providing high-throughput datasets with unprecedented precision and accuracy. The technology for profiling gene expressions and composition (RNA-Seq) has been widely applied in biological/medical research. Appropriate and powerful statistical analysis using RNA-Seq data is essential to the research.

Generating an accurate list of differentially expressed genes is the basis for pathway or gene set enrichment analysis. A gene set with a large number of false positives will compromise these analyses. The parametric approaches discussed here are preferable to nonparametric ones in order to increase the power of detection. However, the false positive rate may be dramatically increased if the assumptions of the model distribution are violated. In Figure 2, we demonstrate the difference between the histogram of 1000 data points simulated from an underlying distribution and the probability mass of an incorrectly fitted model (red curves). The data in the left panel are generated from a Negative Binomial distribution with mean 10 and over-dispersion 0.5 using the *R* function *rnegbin()*, but are modeled by a Poisson distribution with the probability mass:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, \cdots\cdots,$$

where the mean $\lambda = 10$. Those in the right panel are generated from a Poisson distribution with mean 10 using the *R* function *rpois()*, but are modeled by a Negative Binomial distribution with the probability mass:

$$p(k) = \frac{\Gamma\left(k + \phi^{-1}\right)}{\Gamma(k+1)\Gamma\left(\phi^{-1}\right)}\left(\frac{1}{1+\mu\phi}\right)^{\frac{1}{\phi}}\left(\frac{\mu}{\mu+\phi^{-1}}\right)^{k},$$
$$k = 0, 1, 2, \cdots\cdots,$$

where Γ represents the gamma function and the mean $\mu = 10$, the over-dispersion parameter $\phi = 0.5$. The values for the small circles in the fitted models are calculated as the product of 1000 and *p(k)*. The differences in both cases are not negligible, indicating the seriousness of wrong assumptions about the model.

To our knowledge, there is no paper in the literature which studies the efficacy of these methods when the model assumptions do not hold. Given the limitation of small sample sizes in RNA-Seq experiments, robust test procedures which safeguard against the departure of model assumptions are necessary.

Most of the proposed methods produce raw p values for genes based upon the approximate/asymptotic null distribution. This approximation performs well for highly expressed genes but performs poorly for lowly expressed genes. This may create bias during the selection of differentially expressed genes. Some authors simply filter out lowly expressed genes. This is very subjective, and some important genes, which are expressed in one condition but not in another, may be excluded from the result. New testing approaches, which are powerful and effective for both highly and lowly expressed genes, are still needed.

Competing interests

The authors declare no competing interests.

Authors' contributions

ZF, JM and ZW wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgement

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under contract number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States government, or any agency thereof, or the Regents of the University of California.

Author details

[1]Biostatistics Program, School of Public Health, LSU Health Sciences Center, 2020 Gravier Street, 3rd floor, New Orleans, LA 70112, USA. [2]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA94598. [4]Staff Scientist, Group Lead for Genome Analysis, DOE Joint Genome Institute, 2800 Mitchell Dr., MS100, Walnut Creek, CA 94598, USA.

Received: 21 March 2012 Accepted: 12 June 2012
Published: 31 July 2012

References

1. VanGuilder HD, Vrana KE, Freeman WM: **Twenty-five years of quantitative PCR for gene expression analysis.** *Biotechniques* 2008, **44**(5):619–626.
2. Slonim DK, Yanai I: **Getting started in gene expression microarray analysis.** *PLoS Comput Biol* 2009, **5**(10):e1000543.
3. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, *et al*: **Global identification of human transcribed sequences with genome tiling arrays.** *Science* 2004, **306**(5705):2242–2246.
4. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308**(5725):1149–1154.
5. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, *et al*: **Empirical analysis of transcriptional activity in the Arabidopsis genome.** *Science* 2003, **302**(5646):842–846.
6. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26**(10):1135–1145.
7. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities.** *Nat Rev Genet* 2011, **12**(2):87–98.

8. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**(1):57–63.
9. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat Rev Genet* 2011, **12**(10):671–682.
10. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32**(Suppl):490–495.
11. Schulze A, Downward J: **Navigating gene expression using microarrays–a technology review.** *Nat Cell Biol* 2001, **3**(8):E190–E195.
12. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7**(1):55–65.
13. Wilhelm BT, Landry JR: **RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.** *Methods* 2009, **48**(3):249–257.
14. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics* 2002, **18**(4):546–554.
15. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):R106.
16. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq.** *Nat Methods* 2011, **8**(6):469–477.
17. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.** *BMC Bioinforma* 2010, **11**:422.
18. Jiang H, Wong WH: **Statistical inferences for isoform expression in RNA-Seq.** *Bioinformatics* 2009, **25**(8):1026–1032.
19. Richard H, Schulz MH, Sultan M, Nurnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, *et al*: **Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments.** *Nucleic Acids Res* 2010, **38**(10):e112.
20. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**(5):511–515.
21. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18**(9):1509–1517.
22. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
23. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**(5881):1344–1349.
24. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11**(3):R25.
25. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32**(Suppl):496–501.
26. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23**(21):2881–2887.
27. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA: **Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays.** *BMC Genomics* 2009, **10**:221.
28. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**(1):139–140.
29. Zhou YH, Xia K, Wright FA: **A powerful and flexible approach to the analysis of RNA sequence count data.** *Bioinformatics* 2011, **27**(19):2672–2678.
30. Auer PL, Doerge RW: **A two-stage poisson model for testing RNA-Seq data.** *Stat Appl Genet Mol* 2011, **10**(1):26.
31. Agresti A: **Categorical data analysis**. In *Categorical data analysis*. 2nd edition. New York: Wiley; 2002.
32. Storey JD: **A direct approach to false discovery rates.** *J Roy Stat Soc B* 2002, **64**:479–498.
33. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci USA* 2003, **100**(16):9440–9445.
34. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Stat* 1979, **6**(2):65–70.
35. Hochberg Y: **A sharper bonferroni procedure for multiple tests of significance.** *Biometrika* 1988, **75**(4):800–802.
36. Gu K, Ng HKT, Tang ML, Schucany WR: **Testing the ratio of two poisson rates.** *Biometrical J* 2008, **50**(2):283–298.
37. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2008, **9**(2):321–332.
38. Cox DR, Reid N: **Parameter orthogonality and approximate conditional inference.** *J Roy Stat Soc B Met* 1987, **49**(1):1–39.
39. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *J Roy Stat Soc B Met* 1995, **57**(1):289–300.
40. Baggerly KA, Deng L, Morris JS, Aldaz CM: **Overdispersed logistic regression for SAGE: Modelling multiple groups and covariates.** *BMC Bioinforma* 2004, **5**:144.
41. Hoen PAC t, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, Boer JM, van Ommen GJB, den Dunnen JT: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.** *Nucleic Acids Res* 2008, **36**(21):e141.
42. Pham TV, Piersma SR, Warmoes M, Jimenez CR: **On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics.** *Bioinformatics* 2010, **26**(3):363–369.

doi:10.1186/2045-3701-2-26
Cite this article as: Fang *et al.*: Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & Bioscience* 2012 **2**:26.