Communications for Statistical Applications and Methods
2015, Vol. 22, No. 2, 181–199

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.2.181
Print ISSN 2287-7843 / Online ISSN 2383-4757

How are Bayesian and Non-Parametric Methods Doing a Great Job in RNA-Seq Differential Expression Analysis? : A Review

Sunghee Oh[1,a]

[a]Department of Veterinary Medicine, Jeju National University, Korea

Abstract

In a short history, RNA-seq data have established a revolutionary tool to directly decode various scenarios occurring on whole genome-wide expression profiles in regards with differential expression at gene, transcript, isoform, and exon specific quantification, genetic and genomic mutations, and etc. RNA-seq technique has been rapidly replacing arrays with seq-based platform experimental settings by revealing a couple of advantages such as identification of alternative splicing and allelic specific expression. The remarkable characteristics of high-throughput large-scale expression profile in RNA-seq are lied on expression levels of read counts, structure of correlated samples and genes, larger number of genes compared to sample size, different sampling rates, inevitable systematic RNA-seq biases, and etc. In this study, we will comprehensively review how robust Bayesian and non-parametric methods have a better performance than classical statistical approaches by explicitly incorporating such intrinsic RNA-seq specific features with flexible and more appropriate assumptions and distributions in practice.

Keywords: RNA-seq, differential expression, alternative splicing, allelic specific expression, Bayesian and non-parametric methods.

1. Introduction

With the advent of new elegant count based technology referred to as RNA-seq, newly developed RNA-seq specific methods and adaptively implemented ones from microarrays have been proposed and successfully contributed to transcriptome studies (Gerns Storey *et al.*, 2014; Ginsberg *et al.*, 2010; Han and Jiang, 2014; Kim *et al.*, 2012; Kumar *et al.*, 2012; Li *et al.*, 2014; Mills *et al.*, 2013; Nishiu *et al.*, 2002; Satoh *et al.*, 2014; Wang *et al.*, 2013; Warren *et al.*, 2015; Zhang *et al.*, 2014). Representative statistical methods have been popularly utilized to quantify expression levels of mRNA abundance and identify differentially expressed genes between multiple conditions (Anders and Huber, 2010; Anders *et al.*, 2013; Anders *et al.*, 2012; Bar-Joseph *et al.*, 2012; Bi and Davuluri, 2013; Bullard *et al.*, 2010; Glaus *et al.*, 2012; Hardcastle and Kelly, 2010; Lee *et al.*, 2011; Marioni *et al.*, 2008; Niu *et al.*, 2014; Oh *et al.*, 2013; Oshlack *et al.*, 2010; Pollier *et al.*, 2013; Roberts *et al.*, 2011; Robinson *et al.*, 2010; Robinson and Oshlack, 2010; Tarazona *et al.*, 2011; Trapnell *et al.*, 2009; Trapnell *et al.*, 2012; Young *et al.*, 2010).

To date, comparative studies between distinct mRNA samples and groups have become a routine procedure in RNA-seq, nonetheless, each method still has limitations to be further improved

[1] Department of Veterinary Medicine, Jeju National University, 1 Ara 1 Dong, Jeju City, Jeju Do, 690-756, S. Korea. E-mail: sshshoh1105@gmail.com

Published 31 March 2015 / journal homepage: http://csam.or.kr

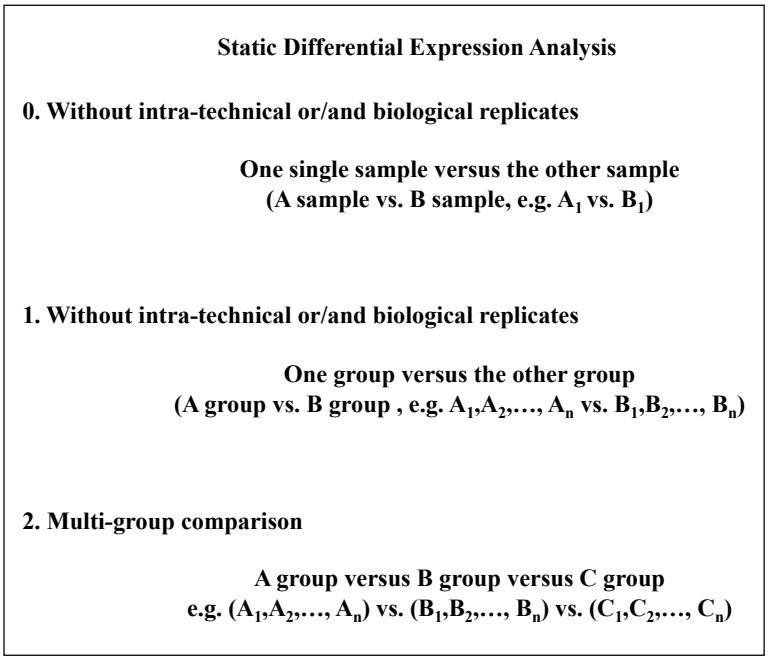

Figure 1: *Static differential expression analysis.*

by explicitly addressing inherent variation and systematic artifacts across samples. Various scenarios in comparative study using gene expression profile are illustrated in Figure 1. As shown in the schematic illustration, all mRNA samples are independently distributed and there is not correlated structure between consecutive samples at all on assumptions in the experimental settings at static differential expression analysis. On the contrary, dynamic methods assume a dependent structure between neighboring samples, for instance, markov assumption such that current state is affected by the right previous state in the variety of time series experimental designs.

When RNA-seq platform has been firstly introduced, pairwise comparative study has been widely conducted to compare two conditions by accounting for count property of expression levels as similarly done in microarrays using classical *t*-test or ranked wilcox test. For example, Fisher exact test has been proposed for this purpose simply to read counts on different mRNA samples in 2×2 contingency table based on hypergeometric distribution (Bullard *et al.*, 2010). The major drawback of this method is limited to the comparison of two groups and it does not explicitly take into account the variability of biological intra-samples lacking the assessment of reproducibility between replicates. And it results in greatly relying upon magnitude of expression on testing in differential expression by demonstrating more differentially expressed genes at higher expression levels. To precisely access variability of replicates and conclude more confident outcomes, the experimental design with the proper number of intra-samples has been recommended in the manner of well-balanced experimental design (Marioni *et al.*, 2008). As a prominent methodology work to precisely incorporate variability of replicates, negative binomial distribution has been next developed in edgeR and DESeq package with a dispersion parameter in their models (Anders and Huber, 2010; Anders *et al.*, 2013; Robinson *et al.*, 2010; Robinson and Oshlack, 2010).

Followed by simple pairwise comparison, multi-group comparison has been commonly performed and both packages have also incorporated the feature enabling to carry out generalized linear models

containing more than two groups and other nuisance factors. Practically, a bottleneck to develop RNA-seq specific methods is insufficient sample size compared to relatively large-size of variables (genes and transcripts). And the calculation of optimal number of intra-replicates and inter-samples with power of detection by controlling FDR has been performed prior to differential expression analysis to derive more reliable statistical testing of comparison. In order to resolve these current issues in RNA-seq methods, Bayesian and non-parametric methods have been proposed and popularly done thus far by demonstrating equivalent at least or better performance in statistical tests compared to existing methods that are on the basis of classical parametric assumptions (Bi and Davuluri, 2013; Hardcastle and Kelly, 2010; Hu *et al.*, 2014; Lee *et al.*, 2011; León-Novelo *et al.*, 2014; Nariai *et al.*, 2013; Nariai *et al.*, 2014; Oh *et al.*, 2013; Ryan *et al.*, 2012; Shen *et al.*, 2012; Shen *et al.*, 2014; Tarazona *et al.*, 2011).

The remainder of manuscript will be discussed in details how Bayesian and non-parametric methods in rigorous template have been adopted and effectively addressed unknown various sources of bias, variability of samples, and other RNA-seq specific natures which have been thoroughly unable to address in the exiting methods.

2. Methods

Profiling high-throughput large-volume of RNA-seq expression data has been a fundamental technique to unveil unknown biological mechanisms in transcriptome studies. RNA-seq has advantageous features that have not been detected in previous platforms in the following, the dynamic range of expression levels, higher quality of samples, and identification of diverse architecture in isoform splicing events and tissue specific allelic imbalance, *etc*. In spite of the considerable advance of next generation sequencing technology, researchers have steadily further addressing various issues to intrinsically arise in RNA-seq specific methodological, experimental, and technical perspectives. In this section, each Bayesian and non-parametric method will be introduced and discussed how it is effectively applied to new RNA-seq expression profile data.

2.1. baySeq

baySeq method has been implemented based on an empirical Bayesian approach in gamma poisson and empirical negative binomial distribution in the very beginning of RNA-seq comparative study. It enabled to increase the accuracy of prediction in differential expression analysis by borrowing information across whole observed expression profiles for genes and samples (Hardcastle and Kelly, 2010). Additionally, it allows us to analyze more complicated experimental designs as a more general statistical testing framework, whereas, the previously existing approaches are all restricted to pairwise comparison and they do not consider the variability induced from intra-samples under the given condition, either. In the evaluation, baySeq demonstrated that it either equivalently performs or outperforms existing methods, over-dispersed logistic, over-dispersed log linear model, DEGSeq, DESeq, and edgeR in both simulated and real data applications in terms of true discovery rates and power of detected calls.

In the methodological rationale, by addressing biological hypotheses in the diversity of experimental designs, empirical Bayesian approach infers the posterior probabilities of each of a set of models that present various combinations of differential expression patterns whether to be differentially expressed or equally expressed for each tuple as non-overlapping sets of samples. The simplest model consists of two possible models for two different conditions, A and B, equally expressed (EE) = $\{A, B\}$ and differentially expressed (DE) = $\{A\}, \{B\}$. Let count-basis expression levels be a set of n samples,

$A = \{A_1, \ldots, A_n\}$ and the observed data for a particular tuple c be $(u_{1c}, \ldots, u_{nc})$ where u_{ic} is the read count for a particular tuple c for sample i. And library size which is related with sequencing depth for each mRNA sample represents by l_i for each tuple, a particular expression pattern,

$$D_c = \{(u_{1c}, \ldots, u_{nc}), (l_1, \ldots, l_n)\}.$$

On the assumption with some model M by the sets, $\{E_1, \ldots, E_m\}$, for the sake of simplicity, if A_i and A_j are from the identical set E_q and have the same parameters of underlying distribution θ_q, $K = \{\theta_1, \ldots, \theta_m\}$ the parameter set of given model M for the expression profile count data, the posterior probability of the model M given the data D_c is given by

$$p(M|D_c) = \frac{p(D_c|M)p(M)}{p(D_c)}, \tag{2.1}$$

$$p(D_c|M) = \int p(D_c|K, M)p(K|M)dK, \tag{2.2}$$

when appropriate intra-replicates are available on data. Negative binomial or over-dispersed poisson distribution is generally adopted for the underlying distribution in differential expression testing on raw read counts or corrected expression levels after normalization. In baySeq, it estimates the over-dispersion of biological replicates by retaining different library sizes for samples, l_i

$$u_{ic} \sim NB\left(\mu_q l_i, \varphi_q\right), \tag{2.3}$$

where $\theta_q = (\mu_q, \varphi_q)$. Since there is no directly matched conjugate prior for this distribution, numerical technique is instead applied by defining an empirical distribution on K and we then estimate $p(D_c|M)$,

$$p(D_c|M) = \int p(D_c|K, M), \tag{2.4}$$

$$p(K|M)dK = \prod_q p(D_{qc}|\theta_q)p(\theta_q)d\theta_q, \tag{2.5}$$

and an empirical distribution on K is derived by examining the entire data set. The major difficulty in this procedure is to estimate degree of dispersion, over- and under-dispersion. Firstly, as the most optimal approach when having a very few number of samples, it assumes a common dispersion that is identical for a tuple across different sets of samples.

In the evaluation of baySeq with others at the constant dispersion, ROC curves demonstrate that baySeq is as good as or outperforms in terms of power of detection on controlled FDR. This method is currently available in R and bioconductor package in the community. And it has advantages in regards to allowance of more complex experimental designs, estimation of dispersion parameter as well as differential expression between conditions, albeit in the case of small sample size. In principle, varying differential expression patterns are given in the estimation of posterior probabilities across multiple conditions (more than two) with biological intra-replicates.

As demonstrated in the paper, baySeq method depicts the real data example for the pairwise comparison of two RDR6 (RNA dependent RNA polymerase 6) knockout samples in the dataset of Illumina sequencing from leaf samples of Arabidopsis thaliana. In the biologically expected hypothesis, it is very well known that RDR6 is required for production of tasRNAs (trans-acting small RNAs). Hence, the central goal of differential expression using baySeq is to identify statistically significant

difference between wild-type and mutant samples with two intra-replicates implicating that baySeq outperforms other competing methods by demonstrating higher rates on true biological findings.

To account for the variability of intra-replicates, negative binomial distribution has been applied by inferring over- or under-dispersion parameter between samples due to different library sizes and sample differences. In the results, baySeq method identified 678 different small RNA sequences that perfectly matched the tasRNA loci, namely, matched nowhere else in the genome. In the comparative outcomes, both edgeR and baySeq detect considerably more tasRNA associated small RNAs than DESeq method, over-dispersed logistic and over-dispersed log-linear approaches. Furthermore, baySeq identifies more tasRNA associated small RNA sequences than edgeR for a given number of selected small RNA sequences. As we demonstrated in this section, baySeq has been introduced as an initial approach when RNA-seq was in its infant. Strikingly, albeit it is the initial model, it allows us to contain biological replicates in two different experimental settings including simple pairwise and multi-group comparison. And also, it presents the posterior probabilities for differential and equal expression pattern, how much likely the given gene is expressed for differential and equal expression, whereas other methods simply do determine whether a gene of interest is either differential expression or equal expression. Yet, more complicated design with multiple factors (*e.g.* age, gender, region, and other profile information for individuals) for each group or time dependent structure are not applicable in the current setting of baySeq. Hence, statistical modeling strategy for more complex comparative studies including time series data is still remained as an open question to be further developed in the community.

2.2. BM-DE

Another approach, BM-DE (a Bayesian method of calling differential expression) models the position-level read counts within a gene by considering position outliers that distributed with un-uniform over positions (Lee *et al.*, 2011). The common approaches to aggregate such positions into a gene level count might be misleading quantification of mRNA abundance. Because, existing methods do not completely take the possibility of variability for position level read counts to be summarized with single gene quantification level. As mentioned in the study, similarly, loci-specific inference approach proposed by Ji and Liu (2010) improved the performance of detection in differential expression by incorporating the feature to borrowing information across loci via hierarchical model even when there are no available intra-replicates. BM-DE followed this approach and conditional on the total count

$$N_{ij}, n_{ij} \sim \text{Bin}\left(N_{ij}, p_{ij}\right), \tag{2.6}$$

fog the positions j, where p_{ij} represents the true proportion of the read count under the condition and relative value to the total read count under both conditions at location j of gene i. In the modeling approach to down-weight p_{ij} for influencing positions in the inference for differential expression,

$$p_{ij}|w_{ij}, \alpha_i, \beta_i \overset{iid}{\sim} \begin{cases} \text{Be}\left(\alpha_i, \beta_i\right), & \text{if } w_{ij} = 1, \\ \text{Be}\left(\frac{1}{2}, \frac{1}{2}\right), & \text{if } w_{ij} = 0, \end{cases} \tag{2.7}$$

when $w_{ij} = 0$, the j^{th} position on the gene is an outlier and prior is forced to assign most probability mass close to 0 or 1. By following the notations as described in Lee *et al.* (2011)

$$\begin{aligned} \eta_i &= \log\left(\alpha_i + \beta_i\right), \\ \xi_i &= \log\left(\frac{\alpha_i}{\beta_i}\right), \end{aligned} \tag{2.8}$$

where ξ_i is the logit of the mean $\alpha_i/(\alpha_i + \beta_i)$ of the beta distribution.

In the set of main parameters (ξ_i, η_i), peculiarly large or small value of ξ_i indicates differential expression and η_i represents varying levels of heterogeneity across genes. Based on the mixture of normal distribution for ξ_i, differential expression is estimated by the notation,

$$\xi_i|\bar{\xi}, s_\xi^2 \sim \pi_0^\lambda N\left(\bar{\xi}, s_\xi^2\right) + \pi_{-1}^\lambda N\left(\bar{\xi} - \delta_{-1}, s_\xi^2\right) + \pi_1^\lambda N\left(\bar{\xi} + \delta_1, s_\xi^2\right), \tag{2.9}$$

where a latent trinary parameter $\lambda_i \in \{0, -1, 1\}$ represents equal, under, and over-expression levels. And it is re-notated by the equation,

$$\xi_i|\lambda_i, \bar{\xi} \sim N\left(\bar{\xi} + \lambda_i\delta_{\lambda_i}, s_\xi^2\right), \quad \Pr\left(\lambda_i = l\right) = \hat{\pi}_l, \quad l = -1, 0, 1. \tag{2.10}$$

In the evaluation, position level Bayesian modeling approach demonstrates more robust estimates by down-weighted outliers on the gene compared to analysis of sequence counts (ASC) and DEGSeq method in synthetic data sets and yeast real data application. As presented in the paper, in the real data application, mRNA samples were collected from yeast Saccharomyces cerevisiae strain BY474 in the comparison of rich growth medium (YEPD medium) and poor growth medium (amino acid starvation). The central purpose of RNA-seq data analysis is to identify genes that are differentially expressed under these two conditions. They first filtered out genes containing greater than 5 positions, remaining, I = 1,089 genes with the range of reads from 1 to 9,334 and from 0 to 14,150 under the distinct conditions, respectively.

Further hierarchical structure with positions from gene levels demonstrates that genes have many positions with non-zero read counts and reads per position are small. Thus, authors presented a yeast experimental example for differential expression at position level counts and eventually estimation of more robust gene expression levels. This framework conclusively showed effective approach by down-weighting outlying observations at position level. Proposed BM-DE method outperformed other methods, in particular, there are position level outliers on the given gene. For the circumstance, the hierarchical modeling approach aims at inferring to position level expression counts other than estimating expression levels at genes by accounting for the extreme values on positions affecting to gene level counts. Nonetheless, the method has the major limitation to be applied for RNA-seq differential expression analysis. BM-DE does not account the variability of intra-biological replicates in the comparison. In order to access reproducibility and variability of intra-replicates, authors remained the part as an important future study that might be straightforwardly extended with the multinomial likelihood in dirichlet prior.

2.3. NOISeq

NOISeq has been proposed in the purpose to precisely account for different sequencing depth across samples on the basis of data-adaptive non-parametric strategy (Tarazona *et al.*, 2011). More specifically, it has been developed by comprehensively investigating the relationship between sequencing depth and differential expression in order to investigate the questionnaire how sequencing depth affects the detection of transcripts and corresponding differential expression, respectively. Basically, NOISeq estimates differential expression at gene levels, although it is possible to straightforwardly extend this method for trasncripts and exon quantification. In particular, the gene expression level in this method is defined by the number of reads or in the library mapping to a gene, that is, the read counts.

Let c_{gj}^i denote the number of read counts for each gene i in the j^{th} sample from different conditions, g = 1 and 2. And s_{gj} denotes the library size computed as the sum of counts c_{gj}^i over all the

genes. For the j^{th} replicate in the experimental condition g^{th}, the corrected expression values by the transformation to control library size bias,

$$x_{gj}^{i} = \frac{c_{gj}^{i} \times 10^{6}}{s_{gj}}. \tag{2.11}$$

Other normalization methods such as RPKM and TMM are also applicable here prior to differential expression analysis (Anders *et al.*, 2013; Bullard *et al.*, 2010; Mortazavi *et al.*, 2008; Robinson and Oshlack, 2010; Wagner *et al.*, 2012). For a particular gene i,

$$M^{i} = \log_2\left(\frac{x_1^j}{x_2^j}\right), \quad \text{log ratio}, \tag{2.12}$$

$$D^{i} = \left|x_1^i - x_2^i\right| = \text{the absolute value of difference}. \tag{2.13}$$

For all of genes, M^i are D^i calculated and a user-defined cutoff threshold value must be established to classify genes whether tested gene is differential expression representing that M and D are very likely to be higher values than noise. Otherwise, it is defined as equal expression. The M and D probability distribution in noise data is computed by contrasting gene counts under the assumption that all samples are collected from the same condition. First of all, M^* and D^* supposedly represent the random variables describing noise distribution. And G^i denotes a random variable to equal to 1 if gene i is differentially expressed between two different conditions. Otherwise, it equals to 0. Therefore, the probability to be differential expression is written by,

$$p\left(G^i = 1 | x_1^i, x_2^j\right) = p\left(G^i = 1 | M^i = m^i, D^i = d^i\right) = P\left(|M^*| < \left|m^i\right|, D^* = d^i\right). \tag{2.14}$$

Accordingly,

$$p\left(G^i = 0 | M^i = m^i, D^i = d^i\right) = 1 - \Pr\left(|M^*| < \left|m^i\right|, D^* = d^i\right) \tag{2.15}$$

and the ratio of two notations as odds values can be also utilized as the testing of differential expression.

NOISeq algorithm computes the probability distribution for M^* and D^* in an empirical way obtaining M and D values for all possible pair of replicates within the same experimental condition for every gene. Those are further used to generate noise distribution by pooling such M and D values with the biological intra replicates under the particular condition, ${}_{J_g}C_2$ times are performed to estimate noise distribution, where J_g is the number of samples in one experimental condition and we assume this method is run sufficiently when two intra-replicates are available. For computing time to run the algorithm, a randomly chosen number from ${}_{J_g}C_2$ can be also applied as well. In contrast, when there are no available replicates, NOISeq method simulates them instead under the assumption that read counts at genes follow a multinomial distribution. In principle, the standard deviation for simulated samples is generated randomly from a uniform distribution in the interval,

$$\left[(pnr - v) \times s_g,\ (pnr + v) \times s_g\right], \tag{2.16}$$

where the parameter *pnr* determines the number of reads of each simulated replicate and is a percentage of the SD, s_g of the available sample g and v is a parameter to represent the variability of SD across samples. Both parameters can be chosen user-specifically and NOISeq allows them to select

the number of replicates to be simulated. Before doing this procedure, exploratory analysis to select a proper number of sample size in power test is recommended to perform in advance to increase confidence of results in differential expression test procedure. In the evaluation of NOISeq method, its better performance is seen when compared to existing methods, edgeR, DESeq, and baySeq on the basis of true positive and negative rates. Other parametric methods such as edgeR demonstrated that large library size data sets produced considerably many false discoveries at low expression levels and/or with small fold changing differences. Thus, parametric approaches are prune to detect more genes when more sequencing reads are used. In summary, as more reads are expressed, as more differential expression calls and noisier data are detected. Improvements of library preparation protocols, sequencing and mapping precision will aid in effectively identifying differential expression in transcriptome architecture.

As introduced in the current section, NOISeq has been proposed as an attractive tool in RNA-seq differential expression analysis by addressing the question how to precisely infer the changes of expression patterns over samples on the basis of more robust non-parametric technique than existing methods. It is fundamentally considering several RNA-seq specific features in the method: sequencing depth that is not warranted to be identical across samples and also, the read length by exploring the relationship among sequencing depth, the distribution of reads, fold change, and differentially expressed genes. Although it is shown in the better performance in terms of true discovery rates compared to other parametric methods, this method is also limited to a static method implicating that dynamic methods taking into account time dependency between samples should be further implemented by researchers.

2.4. NPEBSeq

A novel framework based on non-parametric empirical Bayesian approach (NPEBSeq) has been introduced by Bi and Davaluri (Bi and Davuluri, 2013). In the method, non-parametric nature has been incorporated by empirically estimating from the data without any parametric assumption. As an additional feature of this methodology, it also enables to estimate differential usage of exons as well as gene level analysis. Conclusively, NPEBSeq presented a superior performance than others (baySeq, DESeq, edgeR an NOISeq) based on gold-standard biomarkers in the comparison of percentages on truly differential expression.

In the methodological rationale for single individual mRNA sample without replicates, let x be the number of observed reads mapped to a particularly interesting gene to be tested and r be the expression level of the gene under one condition. X approximately follows a poisson distribution mean $= \lambda = rdl$, where l is the gene length, d is the normalizing constant reflecting the sequencing depth. Given a prior mixing distribution, G with probability density function $g(\lambda)$ on λ. The posterior distribution of λ is

$$g(\lambda)\frac{\lambda^x e^{-\lambda}}{x!}\Big/ h_G(x), \tag{2.17}$$

where

$$h_G(x) = \int \frac{\lambda^x}{x!e^{-\lambda}} dG(\lambda) \tag{2.18}$$

is the G-mixture of poisson.

For the RNA specific property with zero expression level at counts, this method applies a condi-

tioning on $x \geq 1$. To this end, x follows a Q-mixture of zero truncated poisson,

$$\frac{h_G(x)}{1 - h_G(0)} = \int \frac{\lambda^x}{x!(e^\lambda - 1)} dQ(\lambda), \tag{2.19}$$

where

$$dQ(\lambda) = \frac{\left(1 - e^{-\lambda}\right) dG(\lambda)}{\int (1 - e^{-\eta})\, dG(\eta)}.$$

Suppose that n_x denotes the number of genes with exactly x reads in the sample. The conditional non-parametric maximum likelihood estimator $\hat{Q}$ for Q is

$$\hat{Q} = \arg\max \sum_{x \geq 1} n_x \log f_Q(x). \tag{2.20}$$

The posterior distribution of λ is next given by

$$\lambda | x \sim g\left(\hat{\lambda}\right) \frac{\lambda^x e^{-\lambda}}{x!} \Big/ h_{\hat{G}}(x), \tag{2.21}$$

hence empirical bayes estimator for λ is given by

$$\hat{\lambda} = E\left(\hat{\lambda} | x\right) = \frac{(x+1) h_{\hat{G}}(x+1)}{h_{\hat{G}}(x)}. \tag{2.22}$$

The posterior distribution of $\log(d((\lambda_A|x_A)/(\lambda_B|x_B)))$ for different conditions A and B represents log fold change (FC) of expression level of a gene. This can be inferred from the assumption that expected values of log fold change of the majority of genes are zeros,

$$E\left[\log\left(d\frac{\lambda_A|x_A}{\lambda_B|x_B}\right)\right] = 0. \tag{2.23}$$

Therefore, differential expression is tested in NPEBSeq by user-defined cutoff value Δ,

$$\left|\log\left(d\frac{\lambda_A|x_A}{\lambda_B|x_B}\right)\right| > \Delta. \tag{2.24}$$

When biological intra-replicates are available, let c denote the number of biological replicates for one condition and assumptions are based on that

$$\begin{aligned} &x_{ij} \sim \text{POI}(d_j \theta_{ij}), \\ &e_{ij} \sim \text{Gamma}(\lambda_i, \theta), \quad \text{mean} = \lambda_i \quad \text{and} \quad \text{variance} = \lambda_i \theta, \\ &\lambda_i \sim G \text{ such that } g(\lambda) \text{ is the pdf of } G, \end{aligned} \tag{2.25}$$

where x_{ij} is the number of reads for gene i and replicate j and e_{ij} is the expression index. And λ_i is the expression level of gene i under this condition. Q is the scale parameter of Gamma distribution and d_j is the normalizing constant for replicate j. Interestingly, the prior distribution G is estimated by using

the sample that has the largest data depth under each condition. The joint posterior probability of fold change for each gene is given by,

$$\lambda_i, \vec{e}_i|\vec{x}_i \sim g(\lambda_i) \prod_{j=1}^{c} \frac{1}{\Gamma(\lambda_i/\theta)\theta^{\frac{\lambda_i}{\theta}}} \exp_{ij}\left(\frac{\lambda_i}{\theta - 1}\right) \exp\left(-\frac{e_{ij}}{\theta} \frac{(d_j e_{ij})^{x_{ij}} e^{-d_j e_{ij}}}{(x_{ij})!}\right)$$
$$= g(\lambda_i) \prod_{j=1}^{c} \frac{1}{\Gamma(\lambda_i/\theta)\theta^{\frac{\lambda_i}{\theta}}} \exp_{ij}\left(x_{ij} + \frac{\lambda_i}{\theta - 1}\right) \exp\left(-e_{ij}\left(\frac{1}{\theta} + d_j\right) \frac{d_j^{x_{ij}}}{(x_{ij})!}\right). \quad (2.26)$$

NPEBSeq method has been implemented by improving robustness at low expression levels with highly noise values by borrowing information from the gene expression in the entire sample. Another appealing advantage of this method aims at estimation of dispersion by hierarchical Bayesian model and NPEBSeq extended the differential expression to exon levels. Hereby, the major strength in this method is robustness that there are no limited assumptions to be defined previously about the prior distribution and provides the closed form of posterior distribution of fold change.

Let's take a closer look at a real data application hereafter as shown in the paper. In the evaluation of NPEBSeq differential expression method, they employed two real data examples using MAQC (MicroArray Quality Control) dataset and NPEBSeq is compared with alternatives, DESeq, baySeq, and edgeR. The first example is based on MAQC2 Illumina RNAseq data with seven technical replicates of brain reference and those of UHR RNA samples. From 1,000 genes in the original MAQC project, the qRT-PCR gold standard benchmark validated differential patterns, 407 genes are as differential expression and 119 genes as non-differential expression.

In the ROC curves, NPEBSeq obviously outperforms other alternatives in terms of sensitivity and specificity. Interestingly, they also identified differential usage of exons from RNA-seq data to explore the significant effect of the RNAi knockdown of pasilla by RNAseq in the Drosophila melanogaster cell line. The method detected 107 differentially expressed genes under the FDR (false discovery rate) at 0.1 for the comparison of control and knockdown with four intra-replicates. Moreover, NPEBSeq detected differential exon usage for 2,370 counting bins for between condition comparison and for 225 counting bins for within condition comparison at FDR = 0.01. In contrast, alternative DEXSeq method demonstrated much fewer counting bins, 120 as differential expression at FDR = 0.1. In addition, authors compared overlapped common set of two differential expression methods, NPEBSeq and DEXSeq, in both ranked lists of exons. As shown in ROC curves in the paper, NPEBSeq presented better performance than DESeq in terms of true discovery rates both in real data examples and simulated data sets.

2.5. BitSeq

Lastly, we here review another method, bayseian inference of transcripts from sequencing data (BitSeq) (Glaus *et al.*, 2012). In brief, the differential expression is estimated from the posterior samples of expression levels from two or more conditions and all available biological intra-replicates. Samples from the posterior distribution are compared with user-defined threshold value in alteration of expression levels between conditions. Thus, relative expression is represented by Markov Chain Monte Carlo (MCMC) samples from the posterior probability distribution of a generative model of the read counts.

BitSeq analytical pipeline is composed of two main procedures, quantification of transcript expression and estimation of differential expression. The first component in RNA-seq data analysis is to quantify expression level of mRNA abundance of transcripts. For the consistency of previously

mentioned methods, we review the estimation of differential expression part for BitSeq in this study and the quantification of expression levels of transcripts will be further reviewed with another review paper which is currently in the preparation. We assume that the logarithm of transcript expression levels

$$y_m = \log \theta_m, \tag{2.27}$$

where $m = 1, \ldots, M$ and $\theta = (\theta_1, \ldots, \theta_M)$.

Thus, Equation (2.27) represents transcript expression level. For a condition c in the comparison, let R_c be replicate datasets the log expression from replicate r and $y_m^{(cr)}$ be distributed according to a normal distribution with condition mean expression $\mu_m^{(c)}$, normalized by replication specific constant $n^{(cr)}$ and the precision $\lambda_m^{(c)}$,

$$y_m^{(cr)} \sim N\left(\mu_m^{(c)} + n^{(cr)}, \frac{1}{\lambda_m^{(c)}}\right).$$

The conditional mean expression is normally distributed

$$\mu_m^{(c)} \sim N\left(\mu_m^{(0)}, \frac{1}{\lambda_m^{(c)}\lambda_0}\right),$$

where mean $\mu_m^{(0)}$ is empirically calculated from multiple samples and scale parameter $\lambda_m^{(c)}\lambda_0$ for precision. For the conjugate priors and a closed form of posterior probability is given by the following notations,

$$p\left(\mu_m, \lambda_m | y_m\right) = \prod_{c=1}^{C} \text{Gamma}\left(\lambda_m^{(c)} \middle| a_c, \frac{1}{b_c}\right) \times N\left(\mu_m^{(c)} \middle| \frac{\mu_m^{(0)}\lambda_0 + \sum_{r=1}^{R_c}\left(y_m^{(cr)} - n^{(cr)}\right)}{\lambda_0 + R_c}, \frac{1}{\lambda_m^{(c)}\left(\lambda_0 + R_c\right)}\right), \tag{2.28}$$

where

$$a_c = \alpha_G + \frac{R_c}{2}, \quad b_c = \beta_G + \frac{1}{2}\left(\left(\mu_m^{(0)}\right)^2 \lambda_0 + \sum_{r=1}^{R_c}\left(y_m^{(cr)} - n^{(cr)}\right)^2 - \frac{\left(\mu_m^{(0)}\lambda_0 + \sum_{r=1}^{R_c}\left(y_m^{(cr)} - n^{(cr)}\right)\right)^2}{\lambda_0 + R_c}\right).$$

Through the procedure of MCMC, λ_m and μ_m are directly sampled given each pseudo-data vector y_m constructed from the stage 1 MCMC sample. For two comparable conditions, c_1 and c_2, samples of $\mu_m^{(c1)}$ and $\mu_m^{(c2)}$ are used to estimate the probability of expression level o transcript m in condition c_1 being greater than the expression level in c_2. This step is run by the algorithm that counts the fraction of samples to hold the criterion that

$$p\left(\mu_m^{(c_1)} > \mu_m^{(c_2)} \middle| R\right) = \frac{1}{N}\sum_{n=1}^{N} \delta\left(\mu_{m,n}^{(c_1)} > \mu_{m,n}^{(c_2)}\right), \tag{2.29}$$

where $n = 1, \ldots, N$ represents one sample from the above posterior probability for each of N independent pseudo-data vectors.

In the evaluation of BitSeq with other existing methods, its performance is comparable with other competing methods or outperforms DESeq, edgeR, and baySeq in the artificially generated datasets with pre-defined set of differentially expressed transcripts in terms of true discovery rates in ROC

curves. The major advantages of this method aim at the calculation of full posterior distribution, accounting for both technical and biological replicates to compute the posterior distribution of differential expression between conditions to lead fewer false differential expression (DE) calls. In spite of such prominent attractive methodological strategies in recent updates, such as, NPEBSeq and BitSeq, all of introduced methods in the current review have been focused on unified gene level quantification and static method without the sample dependency structure such as markov property. Investigators should upgrade current framework in order to incorporate the important RNA-seq specific advantageous feature, isoform diversity that is defined by multiple ways to combine different exons that have not been able to detect in unified gene level quantification. It is essentially suggesting that those differently alternative splicing events and allelic specific expression across different samples, conditions, and a series of time might play a key role in the clue of aberrant patterns in isoform architecture and mal-functionality behaviors in intricate regulatory mechanism, especially disease related processes such as disease progression with condition for each time.

3. Concluding Remarks

We systematically reviewed various detection tools of differential expression analysis by focusing on Bayesian and non-parametric methods. In the aspect of experimental design setting, that is, the feasibility to more complicated experiments is a critical issue in differential expression analysis. All of described methods allowed us to analyze more than two groups with intra-replicates except BM-DE. In terms of position level quantification other than unified gene level, BM-DE hierarchical modeling approach addressed to the outliers on positions at genes suggesting that single gene level could provide restricted estimates at expression levels when counts are affected by some of influential extreme values on positions. Overall, proposed method in this review outperformed other competing parametric methods in terms of sensitivity and false discovery rates indeed. And more recently, NPEBSeq, NOISeq, and BitSeq have been proposed as outstanding approaches at static comparative study without regards to time points. Although NPEBseq has shown better performance than NOISeq, the method does not explicitly account for RNA-seq specific features, sequencing depth, read distribution, and gene length. Thus, there is no unanimously best method to be suggested under all of the various situations in RNA-seq expression profiles. Furthermore, the sophisticatedly comparative study amongst various methods in a large-scale comparison might need to be further performed in the validation and evaluation of several differential expression methods both in real data application and simulation studies. It will derive more condensed conclusions in the choice of proper RNA-seq specific differential expression methods. Along with such pros and cons, importantly, differential expression analysis has been a commonplace to analyze changes of transcriptional expression profiles over different conditions and samples in several disease mechanisms and organisms in last decades. We highlighted multiple group comparative studies in RNA-seq data that has been more recently proposed and gradually more popularly conducted for a short history in this review article (Anders and Huber, 2010; Anders *et al.*, 2013; Bi and Davuluri, 2013; Bullard *et al.*, 2010; Hardcastle and Kelly, 2010; Lin *et al.*, 2003; Oshlack *et al.*, 2010; Rehrauer *et al.*, 2013; Robinson *et al.*, 2010; Robinson and Oshlack, 2010; Tarazona *et al.*, 2011).

While RNA-seq has a couple of advantageous features such as dynamic rages of expression levels, better quality of sample in terms of reproducibility, identification of isoform architecture and allelic specific expression, it also has RNA-seq specific inherent biases and errors that should be corrected prior to differential expression analysis such as 5' and 3' UTR bias, GC content bias, different sampling rates and sequencing depth across conditions. In addition, the sample size in RNA-seq is much

fewer than the number of variables (genes) (Anders *et al.*, 2012; Beretta *et al.*, 2014; Bernard *et al.*, 2014; Bi and Davuluri, 2013; Bullard *et al.*, 2010; Deng *et al.*, 2011; Glaus *et al.*, 2012; Han and Jiang, 2014; Hiller *et al.*, 2009; Hiller and Wong, 2013; Howard and Heber, 2010; Hu *et al.*, 2014; Jiang and Wong, 2009; Kaur *et al.*, 2012; Kim *et al.*, 2012; Kimes *et al.*, 2014; Kumar *et al.*, 2012; Lee *et al.*, 2011; Leon-Novelo *et al.*, 2014; Lerch *et al.*, 2012; Li *et al.*, 2014; Li *et al.*, 2011; Li and Jiang, 2012; Ma and Zhang, 2013; Marioni *et al.*, 2008; Mezlini *et al.*, 2013; Mills *et al.*, 2013; Mortazavi *et al.*, 2008; Nariai *et al.*, 2013; Nariai *et al.*, 2014; Nicolae *et al.*, 2011; Niu *et al.*, 2014; Oh *et al.*, 2013; Oshlack *et al.*, 2010; Pandey *et al.*, 2013; Patro *et al.*, 2014; Pollier *et al.*, 2013; Rehrauer *et al.*, 2013; Roberts *et al.*, 2011; Robinson and Oshlack, 2010; Ryan *et al.*, 2012; Safikhani *et al.*, 2013; Satoh *et al.*, 2014; Shen *et al.*, 2012; Shen *et al.*, 2014; Shi and Jiang, 2013; Skelly *et al.*, 2011; Suo *et al.*, 2014; Tarazona *et al.*, 2011; Trapnell *et al.*, 2009; Trapnell *et al.*, 2012; Trapnell *et al.*, 2010; Vardhanabhuti *et al.*, 2013; Vitting-Seerup *et al.*, 2014; Wagner *et al.*, 2012; Wang *et al.*, 2010a; Wang *et al.*, 2010b; Wang *et al.*, 2013; Wang *et al.*, 2010c; Wu *et al.*, 2011; Yalamanchili *et al.*, 2014; Young *et al.*, 2010; Zhang *et al.*, 2014; Zhao *et al.*, 2013).

In order to address the artifacts derived from experimental sources, insufficient sample size, and variability of technical and/or intra-biological replicates, more effective Bayesian and non-parametric methods depicted with the methodological details in previous section have been developed in sophisticated manners in this review. They demonstrated equivalent or improved performance when compared to other competing alternative methods both in artificial synthetic and real data application. It implicates that selection of more robust method from the diversity of differential expression tools is important to effectively reduce false discovery rates and provide more reliable biological findings.

4. Closing Remarks

In this review, we learned several empirical Bayesian and non-parametric methods to incorporate RNA-seq specific features by focusing on gene level analytical strategies. In conventional microarray platform, many Bayesian and non-parametric methods have been also popularly introduced. They has been performed analogously with better performance in the aspects of power of detection and false discovery rates in the situation of small samples, large number of probes and genes with noisy sets (Aryee *et al.*, 2009; Gao and Song, 2005; Gerns Storey *et al.*, 2014; Ginsberg *et al.*, 2010; Lin *et al.*, 2003; Nishiu *et al.*, 2002; Stegle *et al.*, 2010; Zhao *et al.*, 2008). More recently, sequencing based platforms have revolutionized transcriptome studies by replacing arrays with RNA-seq. One of major attractive nature of RNA-seq enabled to identify isoform diversity that is created by varying selective ways in exons and to detect allelic imbalance from different transcriptional rates in deep sequencing (Leon-Novelo *et al.*, 2014; Pandey *et al.*, 2013).

In human genes, it is well known that immense amount of gene annotations is associated with these biological phenomena, that is, more than 90% of genes undergo corresponding isoforms. And for an extreme case of gene, a gene has $\sim$ 1,000 isoform splicing events, suggesting that those isoforms with variability might be more closely related with disease and developmental processes as they generate different protein structure and functionalities, respectively (Beretta *et al.*, 2014; Bernard *et al.*, 2014; Deng *et al.*, 2011; Hiller *et al.*, 2009; Hiller and Wong, 2013; Howard and Heber, 2010; Hu *et al.*, 2014; Jiang and Wong, 2009; Katz *et al.*, 2010; Kaur *et al.*, 2012; Kimes *et al.*, 2014; Lerch *et al.*, 2012; Li *et al.*, 2011; Li and Jiang, 2012; Ma and Zhang, 2013; Mezlini *et al.*, 2013; Mills *et al.*, 2013; Nariai *et al.*, 2013; Nariai *et al.*, 2014; Nicolae *et al.*, 2011; Niu *et al.*, 2014; Patro *et al.*, 2014; Rehrauer *et al.*, 2013; Safikhani *et al.*, 2013; Shi and Jiang, 2013; Suo *et al.*, 2014; Trapnell *et al.*, 2010; Vardhanabhuti *et al.*, 2013; Wang *et al.*, 2010b; Wang *et al.*, 2010c; Wu *et al.*, 2011;

Yalamanchili *et al.*, 2014; Zhang *et al.*, 2014).

More importantly, differential expression at gene levels are not guaranteed to be identical cascade patterns either at isoforms or exons, vice versa. NOISeq and NPEBSeq presented exon-specific quantification and analysis as well as gene levels, whereas other methods described in the previous section are all focused on genes. More complex experimental designs have been feasible and gradually showing popularity in clinical application. For example, more complicated experimental designs are widely conducted in the format of multi-series of time course data.

In general, a multi-series of time course data contain biological external conditions (*e.g.* drug treatments) at each time point to address a specific question in disease progression. Taken together, recent studies emphasized that differential expression focusing on gene levels might be a limited approach (Oh *et al.*, 2013; Stegle *et al.*, 2010). And also, as investigators tend to easily overlook the initial experimental design and pre-processing procedures prior to differential expression, as well as more robust and powerful differential expression method, additional critical checking points should be considered in RNA-seq data analysis as given in the following: importance of explorative analysis for diagnosis of samples, proper choice of replicates and samples for well-balanced experimental designs, more deeply sequenced profiles and continuous development of robust statistical methodologies for accurate quantification and differential analysis at genes, transcripts, isoforms, and exons to better understand cellular and molecular complexity (Cumbie *et al.*, 2011; Gatto *et al.*, 2014; Goncalves *et al.*, 2011; Gupta *et al.*, 2012; Hill *et al.*, 2013; Knowles *et al.*, 2013; Kroll *et al.*, 2014; Lin *et al.*, 2012; Martin *et al.*, 2010; Zhang *et al.*, 2012).

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data, *Genome Biology*, **11**, R106.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W. and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor, *Nature Protocols*, **8**, 1765–1786.

Anders, S., Reyes, A. and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data, *Genome Research*, **22**, 2008–2017.

Aryee, M. J., Gutierrez-Pabello, J. A., Kramnik, I., Maiti, T. and Quackenbush, J. (2009). An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation), *BMC Bioinformatics*, **10**, 409.

Bar-Joseph, Z., Gitter, A. and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data, *Nature Reviews Genetics*, **13**, 552–564.

Beretta, S., Bonizzoni, P., Vedova, G. D., Pirola, Y. and Rizzi, R. (2014). Modeling alternative splicing variants from RNA-Seq data with isoform graphs, *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **21**, 16–40.

Bernard, E., Jacob, L., Mairal, J. and Vert, J. P. (2014). Efficient RNA isoform identification and quantification from RNA-Seq data with network flows, *Bioinformatics*, **30**, 2447–2455.

Bi, Y. and Davuluri, R. V. (2013). NPEBseq: Nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data, *BMC Bioinformatics*, **14**, 262.

Bullard, J. H., Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics*, **11**, 94.

Cumbie, J. S., Kimbrel, J. A., Di, Y., Schafer, D. W., Wilhelm, L. J., Fox, S. E., Sullivan, C. M., Curzon, A. D., Carrington, J. C., Mockler, T. C. and Chang, J. H. (2011). GENE-counter: A computational pipeline for the analysis of RNA-Seq data for gene expression differences, *PloS One*, **6**, e25279.

Deng, N., Puetter, A., Zhang, K., Johnson, K., Zhao, Z., Taylor, C., Flemington, E. K. and Zhu, D. (2011). Isoform-level microRNA-155 target prediction using RNA-seq, *Nucleic Acids Research*, **39**, e61.

Gao, X. and Song, P. X. (2005). Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments, *BMC Bioinformatics*, **6**, 186.

Gatto, A., Torroja-Fungairino, C., Mazzarotto, F., Cook, S. A., Barton, P. J., Sanchez-Cabo, F. and Lara-Pezzi, E. (2014). FineSplice, enhanced splice junction detection and quantification: A novel pipeline based on the assessment of diverse RNA-Seq alignment solutions, *Nucleic Acids Research*, **42**, e71.

Gerns Storey, H. L., Richardson, B. A., Singa, B., Naulikha, J., Prindle, V. C., Diaz-Ochoa, V. E., Felgner, P. L., Camerini, D., Horton, H., John-Stewart, G. and Walson, J. L. (2014). Use of principal components analysis and protein microarray to explore the association of HIV-1-specific IgG responses with disease progression, *AIDS Research and Human Retroviruses*, **30**, 37–44.

Ginsberg, S. D., Alldred, M. J., Counts, S. E., Cataldo, A. M., Neve, R. L., Jiang, Y., Wuu, J., Chao, M. V., Mufson, E. J., Nixon, R. A. and Che, S. (2010). Microarray analysis of hippocampal CA1 neurons implicates early endosomal dysfunction during Alzheimer's disease progression, *Biological Psychiatry*, **68**, 885–893.

Glaus, P., Honkela, A. and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation, *Bioinformatics*, **28**, 1721–1728.

Goncalves, A., Tikhonov, A., Brazma, A. and Kapushesky, M. (2011). A pipeline for RNA-seq data processing and quality assessment, *Bioinformatics*, **27**, 867–869.

Gupta, V., Markmann, K., Pedersen, C. N. S., Stougaard, J. and Andersen, S. U. (2012). shortran: A pipeline for small RNA-seq data analysis, *Bioinformatics*, **28**, 2698–2700.

Han, H. and Jiang, X. (2014). Disease biomarker query from RNA-seq data, *Cancer Informatics*, **13**, 81–94.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data, *BMC Bioinformatics*, **11**, 422.

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorsi, B., Su, Y. C. and Yost, H. J. (2013). MMAPPR: Mutation mapping analysis pipeline for pooled RNA-seq, *Genome Research*, **23**, 687–697.

Hiller, D., Jiang, H., Xu, W. and Wong, W. H. (2009). Identifiability of isoform deconvolution from junction arrays and RNA-Seq, *Bioinformatics*, **25**, 3056–3059.

Hiller, D. and Wong, W. H. (2013). Simultaneous isoform discovery and quantification from RNA-seq, *Statistics in Biosciences*, **5**, 100–118.

Howard, B. E. and Heber, S. (2010). Towards reliable isoform quantification using RNA-SEQ data, *BMC Bioinformatics*, **11**, S6.

Hu, Y., Liu, Y., Mao, X., Jia, C., Ferguson, J. F., Xue, C., Reilly, M. P., Li, H. and Li, M. (2014). PennSeq: Accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution, *Nucleic Acids Research*, **42**, e20.

Ji, H. and Liu, X. S. (2010). Analyzing 'omics data using hierarchical models, *Nature Biotechnology*, **28**, 337–340.

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

Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation, *Nature Methods*, **7**, 1009–1015.

Kaur, H., Mao, S., Li, Q., Sameni, M., Krawetz, S. A., Sloane, B. F. and Mattingly, R. R. (2012). RNA-Seq of human breast ductal carcinoma in situ models reveals aldehyde dehydrogenase isoform 5A1 as a novel potential target, *PloS One*, **7**, e50249.

Kim, K. H., Moon, M., Yu, S. B., Mook-Jung, I. and Kim, J. I. (2012). RNA-Seq analysis of frontal cortex and cerebellum from 5XFAD mice at early stage of disease pathology, *Journal of Alzheimer's Disease: JAD*, **29**, 793–808.

Kimes, P. K., Cabanski, C. R., Wilkerson, M. D., Zhao, N., Johnson, A. R., Perou, C. M., Makowski, L., Maher, C. A., Liu, Y., Marron, J. S. and Hayes, D. N. (2014). SigFuge: Single gene clustering of RNA-seq reveals differential isoform usage among cancer samples, *Nucleic Acids Research*, **42**, e113.

Knowles, D. G., Roder, M., Merkel, A. and Guigo, R. (2013). Grape RNA-Seq analysis pipeline environment, *Bioinformatics*, **29**, 614–621.

Kroll, K. W., Kroll, K. W., Mokaram, N. E., Pelletier, A. R., Frankhouser, D. E., Westphal, M. S., Stump, P. A., Stump, C. L., Bundschuh, R., Blachly, J. S. and Yan, P. (2014). Quality control for RNA-seq (QuaCRS): An integrated quality control pipeline, *Cancer Informatics*, **13**, 7–14.

Kumar, R., Lawrence, M. L., Watt, J., Cooksey, A. M., Burgess, S. C. and Nanduri, B. (2012). RNA-seq based transcriptional map of bovine respiratory disease pathogen "Histophilus somni 2336", *PloS One*, **7**, e29435.

Lee, J., Ji, Y., Liang, S., Cai, G. and Muller, P. (2011). On differential gene expression using RNA-Seq data, *Cancer Informatics*, **10**, 205–215.

León-Novelo, L. G., McIntyre, L. M., Fear, J. M. and Graze, R. M. (2014). A flexible Bayesian method for detecting allelic imbalance in RNA-seq data, *BMC Genomics*, **15**, 920.

Lerch, J. K., Kuo, F., Motti, D., Morris, R., Bixby, J. L. and Lemmon, V. P. (2012). Isoform diversity and regulation in peripheral and central neurons revealed through RNA-Seq, *PloS One*, **7**, e30417.

Li, B., Tsoi, L. C., Swindell, W. R., Gudjonsson, J. E., Tejasvi, T., Johnston, A., Ding, J., Stuart, P. E., Xing, X., Kochkodan, J. J., Voorhees, J. J., Kang, H. M., Nair, R. P., Abecasis, G. R. and Elder, J. T. (2014). Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms, *The Journal of Investigative Dermatology*, **134**, 1828–1838.

Li, J. J., Jiang, C. R., Brown, J. B., Huang, H. and Bickel, P. J. (2011). Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation, *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 19867–19872.

Li, W. and Jiang, T. (2012). Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads, *Bioinformatics*, **28**, 2914–2921.

Lin, Y., Reynolds, P. and Feingold, E. (2003). An empirical bayesian method for differential expression studies using one-channel microarray data, *Statistical Applications in Genetics and Molecular Biology*, **2**, 8.

Lin, Z., Puetter, A., Coco, J., Xu, G., Strong, M. J., Wang, X., Fewell, C., Baddoo, M., Taylor, C. and Flemington, E. K. (2012) Detection of murine leukemia virus in the Epstein-Barr virus-positive human B-cell line JY, using a computational RNA-Seq-based exogenous agent detection pipeline, PARSES, *Journal of Virology*, **86**, 2970–2977.

Ma, X. and Zhang, X. (2013). NURD: An implementation of a new method to estimate isoform expression from non-uniform RNA-seq data, *BMC Bioinformatics*, **14**, 220.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays, *Genome Research*, **18**, 1509–1517.

Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. and Wang, Z. (2010). Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads, *BMC Genomics*, **11**, 663.

Mezlini, A. M., Smith, E. J. M., Fiume, M., Buske, O., Savich, G., Shah, S., Aparicion, S., Chiang, D., Goldenberg, A. and Brudno, M. (2013). iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data, *Genome Research*, **23**, 519–529.

Mills, J. D., Nalpathamkalam, T., Jacobs, H. I., Janitz, C., Merico, D., Hu, P. and Janitz, M. (2013). RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism, *Neuroscience Letters*, **536**, 90–95.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, **5**, 621–628.

Nariai, N., Hirose, O., Kojima, K. and Nagasaki, M. (2013). TIGAR: Transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference, *Bioinformatics*, **29**, 2292–2299.

Nariai, N., Kojima, K., Mimori, T., Sato, Y., Kawai, Y., Yamaguchi-Kabata, Y. and Nagasaki, M. (2014). TIGAR2: Sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads, *BMC Genomics*, **15**, S5.

Nicolae, M., Mangul, S., Măndoiu, I. I. and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data, *Algorithms for Molecular Biology: AMB*, **6**, 9.

Nishiu, M., Yanagawa, R., Nakatsuka, S., Yao, M., Tsunoda, T., Nakamura, Y. and Aozasa, K. (2002). Microarray analysis of gene-expression profiles in diffuse large B-cell lymphoma: Identification of genes related to disease progression, *Japanese Journal of Cancer Research: Gann*, **93**, 894–901.

Niu, L., Huang, W., Umbach, D. M. and Li, L. (2014). IUTA: A tool for effectively detecting differential isoform usage from RNA-Seq data, *BMC Genomics*, **15**, 862.

Oh, S., Song, S., Grabowski, G., Zhao, H. and Noonan, J. P. (2013). Time series expression analyses using RNA-seq: A statistical approach, *BioMed Research International*, **2013**, 203681.

Oshlack, A., Robinson, M. D. and Young, M. D. (2010). From RNA-seq reads to differential expression results, *Genome Biology*, **11**, 220.

Pandey, R. V., Franssen, S. U., Futschik, A. and Schlotterer, C. (2013). Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data, *Molecular Ecology Resources*, **13**, 740–745.

Patro, R., Mount, S. M. and Kingsford, C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms, *Nature Biotechnology*, **32**, 462–464.

Pollier, J., Rombauts, S. and Goossens, A. (2013). Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures, *Methods in Molecular Biology*, **1011**, 305–315.

Rehrauer, H., Opitz, L., Tan, G., Sieverling, L. and Schlapbach, R. (2013). Blind spots of quantitative RNA-seq: The limits for assessing abundance, differential expression, and isoform switching, *BMC Bioinformatics*, **14**, 370.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias, *Genome Biology*, **12**, R22.

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*, **26**, 139–140.

Robinson, M. D. and Oshlack, A. A. (2010). Scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biology*, **11**, R25.

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C. and Weinstein, J. N. (2012). SpliceSeq: A resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts, *Bioinformatics*, **28**, 2385–2387.

Safikhani, Z., Sadeghi, M., Pezeshk, H. and Eslahchi, C. (2013). SSP: An interval integer linear programming for de novo transcriptome assembly and isoform discovery of RNA-seq reads, *Genomics*, **102**, 507–514.

Satoh, J., Yamamoto, Y., Asahina, N., Kitano, S. and Kino, Y. (2014). RNA-Seq data mining: Down-regulation of NeuroD6 serves as a possible biomarker for alzheimer's disease brains, *Disease Markers*, **2014**, 123165.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z. X., Zhou, Q., Carstens, R. P. and Xing, Y. (2012). MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data, *Nucleic Acids Research*, **40**, e61.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q. and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, *Proceedings of the National Academy of Sciences of the United States of America*, **111**, E5593–5601.

Shi, Y. and Jiang, H. (2013). rSeqDiff: Detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test, *PloS One*, **8**, e79448.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data, *Genome Research*, **21**, 1728–1737.

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z. and Borgwardt, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series, *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **17**, 355–367.

Suo, C., Calza, S., Salim, A. and Pawitan, Y. (2014). Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data, *Bioinformatics*, **30**, 506–513.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth, *Genome Research*, **21**, 2213–2223.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols*, **7**, 562–578.

Trapnell, C., Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nature Biotechnology*, **28**, 511–515.

Vardhanabhuti, S., Li, M. and Li, H. A. (2013). Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-Sample RNA-Seq Data, *Statistics in Biosciences*, **5**, 119–137.

Vitting-Seerup, K., Porse, B. T., Sandelin, A. and Waage, J. (2014). spliceR: An R package for

classification of alternative splicing and prediction of coding potential from RNA-seq data, *BMC Bioinformatics*, **15**, 81.

Wagner, G. P., Kin, K. and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples, *Theory in Biosciences = Theorie in den Biowissenschaften*, **131**, 281–285.

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010a). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data, *Bioinformatics*, **26**, 136–138.

Wang, L., Xi, Y., Yu, J., Dong, L., Yen, L. and Li, W. (2010b). A statistical method for the detection of alternative splicing using RNA-seq, *PloS One*, **5**, e8529.

Wang, R., Sun, L., Bao, L., Zhang, J., Jiang, Y., Yao, J., Song, L., Feng, J., Liu, S. and Liu, Z. (2013). Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish, *BMC Genomics*, **14**, 929.

Wang, X., Wu, Z. and Zhang, X. (2010c). Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq, *Journal of Bioinformatics and Computational Biology*, **8**, 177–192.

Warren, A.S., Aurrecoechea, C., Brunk, B., Desai, P., Emrich, S., Giraldo-Calderon, G. I., Harb, O., Hix, D., Lawson, D., Machi, D., Mao, C., McClelland, M., Nordberg, E., Shukla, M., Vosshall, L. B., Wattam, A. R., Will, R., Yoo, H. S. and Sobral, B. (2015). RNA-Rocket: An RNA-Seq analysis resource for infectious disease research, *Bioinformatics*, **31**.

Wu, Z., Wang, X. and Zhang, X. (2011). Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq, *Bioinformatics*, **27**, 502–508.

Yalamanchili, H. K., Li, Z., Wang, P., Wong, M. P., Yao, J. and Wang, J. (2014). SpliceNet: Recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples, *Nucleic Acids Research*, **42**, e121.

Young, M. D., Wakefield, M. J., Smyth, G. K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias, *Genome Biology*, **11**, R14.

Zhang, J., Kuo, C. C. and Chen, L. (2014). WemIQ: An accurate and robust isoform quantification method for RNA-seq data, *Bioinformatics*, **30**. The cytochrome P450 genes of channel catfish: Their involvement in disease defense responses as revealed by meta-analysis of RNA-Seq data sets, *Biochim Biophys Acta*, **1840**, 2813–2828.

Zhang, Y., Lameijer, E. W., ’t Hoen, P. A., Ning, Z., Slagboom, P. E. and Ye, K. (2012). PASSion: A pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data, *Bioinformatics*, **28**, 479–486.

Zhao, H., Chan, K. L., Cheng, L. M. and Yan, H. (2008). Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments, *BMC Bioinformatics*, **9**, S9.

Zhao, K., Lu, Z. X., Park, J. W., Zhou, Q. and Xing, Y. (2013). GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data, *Genome Biology*, **14**, R74.

Received February 23, 2015; Revised March 17, 2015; Accepted March 19, 2015