Journal of the Korean Data & Information Science Society
2016, **27**(1), 225–243

http://dx.doi.org/10.7465/jkdi.2016.27.1.225
한국데이터정보과학회지

Beyond gene expression level: How are Bayesian methods doing a great job in quantification of isoform diversity and allelic imbalance?

Sunghee Oh[1] · Chul Soo Kim[2]

[12]Department of Computer Science and Statistics, Jeju National University

Received 14 September 2015, revised 29 October 2015, accepted 29 November 2015

Abstract

Thanks to recent advance of next generation sequencing techniques, RNA-seq enabled to have an unprecedented opportunity to identify transcript variants with isoform diversity and allelic imbalance (Anders *et al.*, 2012) by different transcriptional rates. To date, it is well known that those features might be associated with the aberrant patterns of disease complexity such as tissue (Anders and Huber, 2010; Anders *et al.*, 2012; Nariai *et al.*, 2014) specific differential expression at isoform levels or tissue specific allelic imbalance in mal-functionality of disease processes, etc. Nevertheless, the knowledge of post-transcriptional modification and AI in transcriptomic and genomic areas has been little known in the traditional platforms due to the limitation of technology and insufficient resolution. We here stress the potential of isoform variability and allelic specific expression that are relevant to the abnormality of disease mechanisms in transcriptional genetic regulatory networks. In addition, we systematically review how robust Bayesian approaches in RNA-seq have been developed and utilized in this regard in the field.

Keywords: Allelic imbalance, Bayesian methods, differential expression, gene expression, genetic regulatory network, isoform diversity, post-transcriptional modification, RNA-seq.

1. Introduction

High-throughput gene expression profile data in both array and sequencing have played a key role to address fundamental questions to arise in biomedical research (Anders and Huber, 2010; Anders *et al.*, 2013; Aryee *et al.*, 2009; Bar-Joseph *et al.*, 2012; Bi and Davuluri, 2013; Bullard *et al.*, 2010; Cumbie *et al.*, 2011; Gao and Song, 2005; Hardcastle and Kelly, 2010; Hu *et al.*, 2014; Jiang and Wong, 2009; Lee *et al.*, 2011; Li and Jiang, 2012; Lin *et al.*, 2003; Ma and Zhang, 2013; Marioni *et al.*, 2008; Nariai *et al.*, 2014; Nishiu *et al.*, 2002; Oh *et al.*, 2013; Oshlack *et al.*, 2010; Pollier *et al.*, 2013; Rehrauer *et al.*, 2013; Roberts *et al.*, 2011; Robinson *et al.*, 2010; Robinson and Oshlack, 2010; Shi and Jiang, 2013; Skelly *et al.*,

[1] Assistant professor, Department of Computer Science and Statistics, Jeju National University, Jeju 690-756, Korea.
[2] Corresponding author: Professor. Department of Computer Science and Statistics, Jeju National University, Jeju 690-756, Korea, Email: kimcs@jejunu.ac.kr

2011; Stegle *et al.*, 2010; Suo *et al.*, 2014; Tarazona *et al.*, 2011; Trapnell *et al.*, 2012; Vardhanabhuti *et al.*, 2013; Wang *et al.*, 2013; Wang *et al.*, 2010; Wu *et al.*, 2011b; Zhao *et al.*, 2008). Simultaneous statistical testing based on millions of transcripts and corresponding mRNA samples has made it possible to identify biomarkers that are crucially influencing the alteration of expression levels between disease and normal samples/conditions, classification of sub-groups of particular disease, therapeutic effects on biological external condition such as drug treatments and disease progression over a series of different stages, etc (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). Ultimately, exploring the complexity of disease progressive mechanisms in large-scale of expression profiles cover the entire protocol in the context of biologically hypothetical questionnaire, experimental design, analytical pipeline and corresponding statistical and computational strategy, and validated results with interpretation and take-home messages. For each procedure in the protocol, more effective and reliable methods are required to increase statistical confidence and reduce false discovery rates on biological findings by incorporating systematic variability and biases/errors in generic RNA-seq specific nature (Bullard *et al.*, 2010; Robinson and Oshlack, 2010).

Many statistical methodologies have greatly contributed to the tasks to perform exploratory analysis for diagnosis of samples, identification of changes on expression levels, grouping similar patterns of expression, classification of disease types and detection of system biological network modules and functional pathways. In such an entire pipeline, differential expression analysis is one of major analysis parts enabling identification of disease-specific alteration (or tissue / condition-specific alteration) at gene levels that are directly/indirectly affected by the causality and consequence of disease processes.

Notably, as described in the previous studies, more than 90% of human genes undergo isoform diversity generated by different structures of various exon combinations whose functionality and protein structure also vary, suggesting that characterization at gene levels might lead to a biased and limited conclusion. We will discuss the great potential of post-transcriptional procedure and AI that are more closely related with disease mechanisms as well as gene levels (Aryee *et al.*, 2009; Bar-Joseph *et al.*, 2012; Cumbie *et al.*, 2011; Gao and Song, 2005; Hu *et al.*, 2014; Lee *et al.*, 2011; Marioni *et al.*, 2008; Pandey *et al.*, 2013; Robinson *et al.*, 2010; Skelly *et al.*, 2011; Stegle *et al.*, 2010; Wang *et al.*, 2010; Zhao *et al.*, 2008) by focusing on more enhanced Bayesian methods to detect with flexible assumption and distribution.

Thus, we focus on a selected list of methods based on most updated approaches that outperform existing methods in terms of performance and accuracy, and also naive methods such as cufflinks that have been initially proposed but popularly conducted with comparable performance thus far. Through powerful and rigorous statistical frameworks as proposed in this article, it will aid in investigators detect more accurate isoform architecture and AI specific expression that should be further studied to uncover causality and consequential effect of specific disease of interest in terms of initiation and progression. The remainder will be discussed about both important biological phenomena in post-transcriptional modifications based on statistical and computational methods that have been recently updated and widely applied for the sequencing high-throughput data. The next section will be discussed about improved methods with regards to true discovery rates and power of detection on alternative splicing events and allelic imbalance both in real and synthetic data sets in turn. And we

highlight that entire characterization at spliced structures as well as a single unified gene level will shed a light on comprehensive understanding of unsolved hypothetical questions on transcriptional and genomic studies in the last section. Thus, the identification of post-transcriptional alterations and its better estimation are timely very crucial to fully uncover more precise tissue-specific or external condition-specific biomarkers.

2. Methods

2.1. Identification of tissue specific alternative splicing diversity

The identification of gene expression profiles has been the central role in many of clinical applications and wet experimental laboratories from the conventional array based technology. More recently, with rapidly advanced technologies, ultra high throughput platforms have been proposed accordingly. The improved strategies and methods have a couple of advantages compared to traditional approaches including high resolution, dynamic range of signals on expression levels, better reproducible quality on biological and technical replicates to enable to further explore features such as isoform diversity and allelic specific expression on genetic regulation mechanisms that have not been addressed due to limited resolution and quality on raw data in previous platforms until far. More specifically, isoform variation is structured by one of the post-transcriptional procedures through the variety of selection scheme on given exons in a particular gene, resulting in corresponding various protein structure and functionalities on various spliced isoforms. Thus, analysis at unified gene level might lead to partial conclusion and increment on false discovery rates. Taken together, it is pivotal to explore and characterize unknown underlying biological mechanisms during gene regulation at spliced isoform architectures as well as outer gene level. In order to more precisely decipher post-transcriptional modification in terms of spliced isoforms, a few of methods have been introduced to detect transcript abundance on individual exon and spliced isoform structures statistically and computationally. In the following section, we discuss each method with pros and cons and further demonstrate ability of performance in terms of both real data application and synthetic data.

(1) Tigar2: as RNA-seq expression profiles enable to detect alternatively spliced patterns with the forms of multiple transcript variants in an individual gene, identification of isoform diversity has been performed along with gene level analysis. For the purpose, Tigar2 has been implemented in the environment of Java application as a freely available algorithmic tool (Nariai, *et al.*, 2013; Nariai, *et al.*, 2014). The major strength of this method is the sensitivity and robustness on the quantification of isoform abundance, in particular, it demonstrates higher accuracy of measuring isoform expression levels when varying read length (> 250 base pair). On the contrary, other existing methods sensitively responded to the read length and they presented poor performance in both single and paired-end experiment based on the root mean squared errors of the estimated abundances (log of FRKMs) compared to the true gene expression levels. Thus, the evaluation of Tigar2 is conducted by the comparison of other existing quantification methods in fixed read length (single and paired) and also differently variable read length settings and it presented a better performance compared to other competing alternative methods (Trapnell, *et al.*, 2009; Trapnell, *et al.*, 2012; Trapnell, *et al.*, 2010). In principle, the methodological scheme is on the basis of variational bayesian inference with expectation and maximization. Let θ be a model parameter that represents

transcript isoform abundances, and we assume that Z_{nt} represents an indicator variable. It equals to 1 if read n generated from transcript isoform n, zero otherwise. And R_n^1 and R_n^2 represent nucleotide sequence of the first and second pair of read n, respectively. The joint probability equals to the product of conditional probabilities as given by,

$$p(\theta, Z_{nt}, R_n^1, R_n^2) = p(\theta)p(Z_{nt}|\theta)p(R_n^1, R_n^2|Z_{nt}), \tag{2.1}$$

$p(\theta)$ is represented by dirichlet prior distribution, $p(\theta) = \frac{1}{C}\prod_{t=0}^{T}\theta_t^{\alpha_t - 1}$, where $\alpha_t > 0$ is a hyperparameter, c is constant, T is the number of transcript isoforms, such that $\sum_{t=0}^{T}\theta_t = 1$. θ_0 represents the noise isoform abundance (reads that are not generated from any known isoform are assigned).

$$p(Z_{nt}|\theta) = p(T_n|\theta)p(F_n|T_n)p(S_n|T_n, F_n)p(O_n|T_n)p(A_n^1, A_n^2|T_n, F_n, S_n, O_n), \tag{2.2}$$

where $T_n, F_n, S_n, O_n, A_n^1$ and A_n^2 represent in turn, the transcript isoform choice, fragment size, read start position, orientation, and alignment state of the first pair and second pair of read n. $p(T_n|\theta)$ represents the probability of read n generated from transcript isoform T_n given a parameter vector. A fragment size variable as F_n is included in the Tigar2 model. The conditional probability given $T_n = t_n$ s computed by the notation,

$$p(F_n = f_n|T_n = t_n) = \frac{dF(f_n)}{\sum_{x=1}^{l_t} dF(x)}, \tag{2.3}$$

where l_t represents the length of transcript isoform n, and dF(x) denotes the global fragment size distribution that follows Normal(μ_F, σ_F^2). $p(O_t|T_n)$ represents the probability of the orientation of read n given the transcript isoform choice a strand specific protocol. To account for $p(O_t = 0|T_n = t) = 1$ and $p(O_t = 0|T_n = t) = 0$, and next, $p(A_n^1, A_n^2|T_n, F_n, S_n, O_n)$ represents the probability of the alignment state of read n given the transcript isoform choice, fragment size, start position, and orientation of read n. $p(R_n^1, R_n^2|Z_{nt} = 1)$ is represented by the conditional probability of sequence of the first and second pair or read n given $Z_{nt} = 1$. That is,

$$p(R_n^1, R_n^2|Z_{nt}=1)=\prod_{x=1}^{x^1} emit(r^1[x], q^1[x], c^1[x], a^1[x])\prod_{x=1}^{x^2} emit(r^2[x], q^2[x], c^2[x], a^2[x]), \tag{2.4}$$

where the first term represents the emission probability of nucleotide characters of the first pair of read n, decomposed of nucleotide character $(r^1[x])$, base call quality score $(q^1[x])$, nucleotide character of the corresponding reference sequence $(c^1[x])$, the alignment state of the first pair of read n at position $xa^1[x]$. Likewise, $emit(r^2[x], q^2[x], c^2[x], a^2[x])$ is computed for the second pair. In order to explicitly consider the varying read length distribution,

$$p(L_n = \text{length}(R_n)|F_n = f_n) = \frac{d_R(\text{length}(R_n))}{\sum_{x=1}^{f_n} d_R(x)}, \tag{2.5}$$

where length (R_n) represents the global read length distribution. By denoting with smooth functions in a non-parametric strategy with M equally spaced Gaussian kernels as basis

functions,

$$d_R(x) = \frac{g(x)}{\sum\limits_{x'=1}^{\max(L)} g(x')}, \text{ where } g(x) = \sum_{i=1}^{M} a_i m_i(x), \tag{2.6}$$

a_i represents the coefficient parameters and $m_i(x)$ follows $N(\mu_i, \sigma^2)$ through the sophisticated paradigm of variational bayesian inference approach, latent variables to represent true alignments of reads as well as model parameters (transcript isoform abundances) are computed by the posterior distribution.

Dirichlet prior: $\theta \sim D(\alpha_0, \cdots, \alpha_0),\ \alpha_\theta > 0$, In the prior, $\alpha_0 = 0.001, 0.01, 0.1$ or 1.0 has been considered. In variational bayesian expectation (VBE) step, the expected number of reads that are mapping back to the transcript isoform is obtained by $\widehat{r}_t = \sum_n E_z[Z_{nt} = 1]$. In variable bayesian maximization step (VBM), the expected abundance of transcript isoform t is computed by.

$$E_\theta[\widehat{\theta}_t] = \frac{\widehat{\alpha}_t}{(\sum\limits_t, \widehat{\alpha}_{t'})} \text{ where } \widehat{\alpha}_t = \alpha_0 \widehat{r}_t. \tag{2.7}$$

This improved Tigar2 has been evaluated with other alternative methods in both simulated and real data applications containing technical replicates under the identical biological conditions and variable read lengths. It demonstrates that Tigar2 outperforms other competing methods, Tigar1, RSEM, and Cufflinks in single and paired end data, more specifically, at read length longer than 250 bp (Li and Dewey, 2011; Nariai *et al.*, 2013; Trapnell *et al.*, 2009; Trapnell *et al.*, 2012; Trapnell *et al.*, 2010). Tigar2 is an updated version of Tigar1 additionally featured in loading a BAM file into a genome browser such as Integrative Genomics Viewer. And current version of method has been implemented with improved performance in sensitivity and accuracy of quantification of isoform expression levels from initial Tigar1. More specifically, (Nariai *et al.*, 2014) evaluated human HeLa cell line samples that have obtained 4.25 million reads in single-end sequencing as a real data application. From the data, it demonstrated a right skewed distribution of variable read lengths in the range of ∼50 to ∼300, suggesting that variable length is one of crucial parameters to model a quantification method and evaluate sensitivity and accuracy on quantification of isoform diversity.

(2) SplikeTrap: SpliceTrap has been proposed to quantify transcript variant isoforms on alternative splicing structure. It is based on the independent bayesian inference measuring expression level estimation of each exon. This approach demonstrated better performance than other methods in terms of accuracy, robustness, and reliability in quantifying exon-inclusion ratios. In principle, state of art SpliceTrap tool has been developed to precisely quantify exon-inclusion levels in the format of exon-trio database design referred to as TXdb (Wu, *et al.*, 2011). All known transcripts encoded by each human gene are annotated by the common RefSeq and EST based alternative splicing (AS) database dbASE. In addition, all possible exon-skipping event in AS is identified by subdividing each transcript in to exon trios by sliding a 3-exon window along the transcript. For the quantification of exon-skipping event (cassette exon; CA), some of key measurement metrics are used, the flanking exons (e_1 and e_3) are constitute exons and the middle exon is categorized in to cassette exon and

also an inclusion isoform (f_1) with all three exons and a skipping isoform (f_2) comprising the two flanking exons only.

Specifically, the first and last exons for every transcript are discarded and the transcriptomic variability is primarily due to alternative transcription intiation, or poly-adenylation, other than AS event per se. By using mapping database, TXdb, every exon is independently estimated for its AS level and SpliceTrap has made it able to detect other types of AS structures such as AA: alternative 3' splice site, AD: alternative 5' splice site, and IR: Intron retention. Another advantage of this method is able to capture complex AS events involving two or more exon trios/duos. In a paired-end RNA-seq experiment, a fragment on FPKM is defined as a sequence segment encompassed between the first and last nucleotides of a read-pair. In the methodological framework of SpliceTrap for each trio/duo, the positions of the mapped fragments follow a uniform distribution and corresponding their sizes follow approximately normal distribution. A particular fragment j is denoted as a vector $r_j : (b_j, s_j)$, where b_j and s_j represent the beginning position and size of the fragment, respectively. And also, for exon trio (or exon duo), the set of all possible isoforms are defined as $F = \{f1, f2\}$, where f1 represents an inclusion (or extended) isoform, and f_2 is a skipping (or shortened) isoform. The lengths and the relative expression levels of these isoforms are $L = \{L1, L2\}$ and $E = \{e1, e2\}$. Taken together, the probability of observing an isoform i given the expression level is denoted as,

$$p(f_i|E) = \frac{e_i L_i}{e_1 L_1 + e_2 L_2}, \tag{2.8}$$

supposedly, m represents the number of fragments $R = \{r_j, j = 1, 2, \cdots, m\}$ that can be mapped to F. For given each fragment, $r_j(b_j, s_j)$, $b_j \prod s_j$, the probability of observing r_j given on isoform f_i is

$$p(r_j|f_i, E) = p(b_j|f_i, E), p(s_j|f_i, E) = p(b_j|f_i, E)p(s_i) = \frac{1}{l_i} p(s_j), \tag{2.9}$$

where l_i is the effective length of $f_i(L_i - S_i + 1)$ and $p(s_j)$ is the probability of observing a fragment size s_j in the experiment. For all of isoforms in F, $p(r_j|E)$) equals to

$$\sum_{f_i \in F} p(r_j|f_i, E)p(f_i|E) = \sum_{f_i \in F} \frac{1}{l_i} p(s_j) \frac{e_i L_i}{e_1 L_1 + e_2 L_2} \text{ and } p(R|E) \prod_{r_j \in F} p(f_i|E). \tag{2.10}$$

Finally, a bayesian posterior function is written as

$$p(E|R) \propto \prod_{r_j \in F} p(r_j|E)p(E). \tag{2.11}$$

Thus, SpliceTrap is designed to quantify local alternative splicing activity and exon-inclusion ratios. In order to validate the performance of this method, it is evaluated systematically with other competing methods, RPKM, Cufflinks, Scripture, and MLE in both synthetic and real data application, SpliceTrap presented highly accurate and reliable quantification levels and consistent robustness. In the evaluation, the method is compared in regards to correlation coefficient and mean absolute error in 36 and 75 nucleotide base pair.

Conclusively, SpliceTrap demonstrates more reliability and reproducibility when compared to Cufflinks and Scripture that can also be used to quantify local AS events, albeit with lower accuracy. Here, it is worth noting that SpliceTrap is manifested to quantify various local AS events and exon-inclusion levels for full transcript expression levels. In further details, it was specifically designed to accurately quantify alternative splicing at the single exon level by identifying exon trios/duos instead of full transcripts based on bayesian modeling approach. In order to reduce false discovery rates, various set of cutoff threshold values has been examined. Although SpliceTrap is specifically designed to detect single cassette exons, it can also detect different splicing patterns such as AA and AD described earlier. Similar to Tigar2, authors evaluated the quantification methods using Human Hela cell lines with more than 60 million and 36 nt paired end reads in RNA-Seq real data application. From the data, proposed method presented higher performance in terms of cassette exon discovery rate and specificity on detection of various splicing events, whereas interestingly, SpliceTrap showed less sensitivity on constitute exon discovery rates on varying inclusion ratios compared to others.

(3) BASIS: next, we review a Bayesian analysis of splicing isoforms (BASIS) to estimate the differential expression level of each transcript isoform between two conditions (Zheng and Chen, 2009). In the method, a latent variable is used to predict direct statistical selection of differentially expressed isoforms. Model parameters are inferred based on ergodic Markov chain generated by Gibbs sampler technique. And BASIS has the capability to borrow information across different positions in the content of within-genes and between-genes handling with different sequence depth coverage. In this review, we skip the part of tiling array data since we focus on RNA-seq specific methodology. In BASIS, for each gene, read coverage over each position i that appears in at least one transcript isoform of gene g is modeled with

$$\Delta y_{gi} = \sum \Delta\beta_{gi} X_{gij} + \Delta\epsilon_{gi}, \text{where } \Delta_{gi} \text{ denotes the average difference between two conditions for position } i \text{ of gene } g(\Delta y_{gi} = y^1_{gi} - y^2_{gi}). \tag{2.12}$$

$\Delta\beta_{gi}$ is the expression difference between two conditions. For the j-th transcript isoform of gene g, X_{gij} represents the binary indicator of whether position i belongs to isoform j's exon region, and $\Delta\epsilon_{gi}$ is the error term for position i of gene g. It is conceptually focused on nucleotide positions appearing in at least one transcript isoform. For each position i that appears in at least one transcript isoform of gene g, the read coverage difference under two distinct conditions is denoted by the linear model,

$$\Delta y_{gi} = \sum \Delta\beta_{gi} x_{gij} + \Delta\epsilon_{gi}, \text{where } \Delta y_{gi} = y^1_{gi} - y^2_{gi} \text{ and } \Delta\beta_{gij} \text{ represents the expression difference between two conditions for j-th transcript isoform for j-th transcript isoformof gene g}(\Delta y_{gi} = y^1_{gi} - y^2_{gi}). \tag{2.13}$$

And x_{gij} is the binary indicator of whether position i belongs to isoform j's exon region, $\Delta\epsilon_{gi}$ is the error term. G represents the total number of genes and n_g is the total number of positions for gene g. j ranges from 1 to s_g where s_g is the total number of transcript isoforms for gene g. The total $\Delta\epsilon_{gi}$'s ($g = 1, \cdots, G$ and $i = 1, \cdots, n_g$) are decomposed into

100 bins. Thus, proposed hierarchical bayesian approach is written as

$$\Delta y_g|\Delta\beta_g, \sum_g \sim N_{n_g}(X_g\Delta\beta_g, \sum_g, g = 1, \cdots, G, \sum_g \equiv diag(\prod_{g1}, \cdots, \prod_{gn_g})$$

and $\prod_{gi} = \delta_m$ if position i of gene $g \in bin\ m, \delta_m \sim IG(\nu/2, \nu\lambda/2), m = 1. \cdots, 100.$ (2.14)

$$\Delta\beta_g|\gamma_g \sim N_{s_g}(0, R_g), \text{ and } R_g \equiv daig(k_{g_1}, \cdots, k_{gs_g}),$$
$$\text{where } k_{g_j} = \gamma_{g_j} \text{ if } \gamma_{g_j} = 0 \text{ and } k_{g_j} = \psi_{g_j} \text{ if } \gamma_{g_j} = 1 \tag{2.15}$$

$$f(\gamma_g) = \prod_{j=1}^{s_g} p^{\gamma_{gj}}(1-p)^{1-\gamma_{gj}}, \text{ where } \Delta y_g, \Delta\beta_g \text{ and } X_g \text{are identical as}$$
$$\text{described in the previous equations. And } r_g \text{ is a latent variable.} \tag{2.16}$$

N_{n_g} and N_{s_g} stand for multivariate normal distribution and IG represents the inverse gamma distribution. Given the isoform amount differences ($\Delta\beta_g$) and position arrangements (X_g), read coverage differences (ΔY_g) follow a multivariate normal distribution, MVN($X_g\Delta\beta_g, \sum_g$), where if a position is assigned to bin m, the variance of the coverage difference is $\delta_m \cdot \gamma_{gj}$ is an indicator to represent whether j-th isoform is differentially expressed or not, such that when $\gamma_{gj} = 0, \Delta\beta_{gj} \sim N(0, \psi_{gj})$. In proposed prior distribution for parameters $(\Delta\beta, \delta, \gamma)$, there are hyperparameters $(r, \psi, \nu, \lambda, p)$ and Gibbs sampler techniques are made use of generating Markov Chain and posterior probabilities of $\Delta\beta, \delta$ and γ are in turn estimated from the chain. First of all, the variance parameter $\delta_m^{[0]}$ is initialized to be the mean of intensity summation $(y^1 + y^2)$. For the positions in bin m, $r_m^{[0]}$ is also initialized as $(1, \cdots, 1)^T$. The Gibbs sampler at the k-th iteration is done in the following procedures,

(I) Perform sampling procedure of the isoform amount differences, $\Delta\beta_g^{[k]}(g = 1, \cdots, G)$ from the conditional posterior distribution

$$\Delta\beta_g^{[k]} \sim f(\Delta\beta_g^{[k]}|\Delta Y_g, \delta^{[k-1]}, \gamma_g^{[k-1]} = N_{s_g}\left(AX_g^T\left(\sum\nolimits_g^{[k-1]}\right)^{-1}\Delta Y_g, A\right),$$
$$\text{where } A = \left(X_g^T\left(\sum\nolimits_g^{[k-1]}\right)^{-1}|X_g + \left(R_g^{[k-1]}\right)^{-1}\right)^{-1} \tag{2.17}$$

(II) Perform sampling $\delta_m^{[k]}, m = 1, \cdots, 100$ to represent the variance for positions in bin m, from the conditional posterior distribution,

$$\delta_m^{[k]} \sim f(\delta_m^{[k]}|\Delta Y_m, \Delta\beta_m^{[k]}, \Delta\gamma^{[k-1]}) = IG\left(\frac{\nu + q_m}{2}, \frac{\nu\lambda + (\Delta Y_m - X_m\Delta\beta_m^{[k]})^T(\Delta Y_m - X_m\Delta\beta_m^{[k]})}{2}\right),$$

where are for positions falling in bin m, represents the number of positions in bin m assuming that the positions in bin m may be from different genes. (2.18)

(III) Perform sampling $\gamma_{gj}^{[k]}, g = 1, \cdots, g$ and $j = 1, \cdots, s_g$, indicator variable of whether the j-th isoform should be declared as differentially expressed from the conditional posterior distribution, $\gamma_{gj}^{[k]} \sim f(\gamma_{gj}^{[k]} | \Delta Y, \Delta\beta_g^{[k]}, \Delta^{[k]}, \gamma_{gj}^{[k]})$,

$$Pr(\gamma_{gj}^{[k]} = 1 | \Delta Y, \Delta\beta_g^{[k]}, \Delta^{[k]}, \gamma_{gj}^{[k]} = \frac{f\left(\Delta\beta_g^{[k]} | \gamma_{gi}^{[k]}, \gamma_{gj}^{[k]} = 1\right) p}{f\left(\Delta\beta_g^{[k]} | \gamma_{gi}^{[k]}, \gamma_{gj}^{[k]} = 1\right) p + f\left(\Delta\beta_g^{[k]} | \gamma_{gi}^{[k]}, \gamma_{gj}^{[k]} = 0\right)(1-p)}$$

$$\text{where } \gamma_{gj}^{[k]} = \left(\gamma_1^{[k]}, \cdots, \gamma_{j-1}^{[k]}, \gamma_{j+1}^{[k-1]}, \cdots, \gamma_{sg}^{[k-1]}\right)^T \tag{2.19}$$

In the evaluation of proposed method using RNA-Seq real data application, authors employed mouse brain, liver, and muscle Solexa high-throughput sequencing data. In the quantification of isoform diversity, non-redundant transcript isoform of mouse genes were downloaded from ASTD and Ensembl database. Transcript abundances are estimated for adult mouse brain, liver, and muscle tissue samples. And for each group, there exist two biological individual replicates and uniquely mapped sequence reads from two replicates are all pooled together and mapped to gene annotations. Thus, the known or predicted mouse transcript isoform splicing patterns are obtained from those databases. From the results, pairwise comparisons have been conducted between brain and liver, having that 35,715 transcripts are differentially expressed and around 21,188 transcripts are up-regulated in brain, and the remaining transcripts are up-regulated in liver, respectively. And importantly, around 7,699 genes have more than one differentially expressed transcript isoform from corresponding 10,771 genes

Likewise, in the pairwise comparison between brain and muscle, 34,126 transcripts belonging to 10,554 genes are differentially expressed. Of which 19,851 of the transcripts are up-regulated in brain and others are up-regulated in muscle. And interestingly, 5,498 of 7,392 differentially expressed genes presented at least one down-regulated isoform in brain versus muscle. In the validation of BASIS using RNA-seq data for brain, liver, muscle with two biological individual replicates and simulated sets, the comparison between proposed BASIS and least squares fit is made when the total false-positive rate is controlled at 0.005 and it shows a remarkable improvement on the basis of power test by representing overall more two times better performance between two.

Notably, BASIS method is further evaluated in terms of RT-PCR validation technique by thoroughly examining genes whose isoforms show differential expression patterns tissue-specifically between two conditions. As BASIS infers the differential expression levels by jointly taking into account all positions targeting the individual same gene, the major strength of this method is to aim at considering sequence read coverage at each position that may lead to a family of splice variants instead of one single isoform. In general, the majority of transcript isoforms do not contain any isoform specific sequence positions or isoform specific exon-exon junctions. BASIS address this issue by allocating the intensity of sequence read coverage to multiple transcript isoforms and integrating multiple sequence read coverage values for the same gene. Furthermore, this method has been evidently improved in terms of signal to noise estimate by utilizing shared information across every sequencing read. This approach has been solved the common issue, the large p and small n to arise in high-throughput large scale of expression profiles by incorporating flexible statistical inference compared to traditional least squares fit. Interestingly, the latent γ variable

is involved with variable selection scheme as only small portion of the transcript isoforms is expressed under multiple biological conditions. And it is used to identify the corresponding transcript isoforms of interest and lead to an interpretable model.

(4) RSEM: as RSEM and Cufflinks are compared in the previous section on Tigar2, both methods are also discussed miscellaneously as an appendix in current section. RSEM method has been implemented in an user-friendly software package for both gene and isoform abundances from single and paired end RNA-Seq data with or without reference genome. When a reference genome of interest is not available on mapping and assembly, combination with a de novo transcriptome assembler is utilized. Major strength of this method is that it has superior or comparable performance to others that rely on a reference genome. And this is also an updated version of quantification of splicing events (Li *et al.*, 2010) and has been newly implemented by extending following features: (1) Firstly, paired end reads are modeled in current version using a pair of observed random variables, R^1 and R^2. Accordingly, in the case of single end reads, R^2 is treated as a latent random variable. (2) Secondly, the length of the fragment from which a read or pair of reads is derived is also modeled by the latent random variable F. The distribution of F is specified by a global fragment length distribution, $\lambda_F(x)$ in the following equation,

$$P(F = x|G = i) = \lambda_F(x) \sum_{x^1}^{l_i} \lambda_F(x')^{-1} \text{ where } l_i \text{ is the length of transcript } i. \tag{2.20}$$

As presented in the given notation, $\lambda_F(x)$ is truncated and normalized given a fragment is derived from a specific transcript of finite length. (3) Thirdly, updated method allows various read lengths by representing the observed random variable L (or L^1 and L^2 for paired end). And (4) Finally, in order to account for quality of reads, authors included quality score parameter as a random variable Q. In the quantification of splicing events, this approach does not specify distribution of Q random variables because they are all observed without any dependence upon other variables. RSEM's main rationale is on the basis of computing the Maximum Likelihood values of the parameters, θ of model contained by introduced parameters. θ_i represents the probability that a given fragment is derived from transcript i and when the noise transcript is defined with θ_0 and reads have no alignments. After estimation, the θ values are converted into transcript fractions,

$$\begin{aligned}
&\frac{\theta_i/l_i^{'}}{\sum\limits_{j\neq 0} \theta_i/l_i^{'}}, \text{ where } l_i^{'} \text{ is the effective length of transcript } i \text{ given by} \\
&\sum_{x\leq l_i} \lambda_F(x)(l_i - x + 1) \text{ for poly(A) transcripts and,} \\
&\sum_{x\leq l_i + l_A} \lambda_F(x)\min(l_i + l_A - x + 1, l_i) \text{ for poly(A)+ transcripts,} \\
&\qquad \text{where } l_A \text{ is the length of a poly(A) tail.}
\end{aligned} \tag{2.21}$$

RSEM also obtain posterior mean of estimates and corresponding 95% confidence intervals (CIs) with a two-stage sampling process by Gibbs Samplers.

(5) Cufflinks: the primary goal of Cufflinks is to accurately estimate transcript abundance by accurately identifying which isoform variation from a gene produces each read. It relies upon known all splice variants (isoforms) of the particular gene of interest. Basically, Cufflinks assembles individual transcripts from RNA-Seq reads that have been aligned against a reference genome. Similar to other methods discussed in the previous sections, Cufflinks quantify splicing structure of the gene. It takes into account the fact that genes undergo multiple alternative splicing events usually reconstructing gene models on quantification. In the algorithm, it produces a parsimonious transcriptome assembly data and a few full-length transcript fragments to explain all the possible splicing events for input data. In principle, transcript abundances are estimated based on a generative statistical model technique of RNA-Seq samples. For the sake of simplicity and convenience of proposed model, abundances of non-overlapping transcripts in disjoint genomic loci are computed independently. Model parameters are the non-negative abundances, ρ_t. Thus, the effective length of a transcript is defined by,

$$\bar{l(t)} = \sum_{i=1}^{l(t)} F(i)(l(t) - i + 1)$$

where $l(t)$ is the length of a transcript and F is the distribution of fragment. (2.22)

Therefore, likelihood function is given by,

$$L(\rho|R) = \prod_{r \in R} \prod_{t \in T} \frac{\rho_t \bar{l(t)}}{\sum\limits_{u \in T} \rho_u \bar{l(u)}} \left(\frac{F(l_t(r))}{l(t) - l_t(r) + 1} \right)$$

where R and T represent all fragment alignments and Transcripts, respectively. (2.23)

And $l_t(r)$ represents the implied length of a fragment determined by a pair of reads assuming, it originated from transcript t. This likelihood function has a unique maximum via numerical optimization procedure as it is the likelihood function for a linear model and it is identifiable. More specifically, maximum a posteriori estimates are also computed by a Bayesian inference procedure based on importance sampling from a posterior distribution. In summary, the proposed distribution is a multivariate normal with mean given by the maximum likelihood estimate a variance-covariance matrix given by the inverse of the observed Fisher Information matrix. Thus, the maximum a posteriori estimates are eventually used for differential expression testing.

2.2. Identification of tissue specific allelic imbalance

Now, we comprehensively review methods to identify allelic specific expression and allelic imbalance (AI) as the product of genetic differences in gene regulation (Leon-Novelo *et al.*, 2014; Ng *et al.*, 2014; Pandey *et al.*, 2013). This is primarily induced by the circumstances when regulatory processes result in different steady-state transcriptional rates for two alleles within a single biological individual. The goal of firstly introduced approach proposed by Leon-Novelo LG *et al* (Leon-Novelo *et al.*, 2014) is to correct the bias in estimation of AI derived from multiple factors, genome ambiguity, reference quality, the mapping algorithm, and biases in sequencing procedures.

In the methodological tasks, the underlying mechanisms are based on a flexible Bayesian model for analysis of AI to account for bias and implementation without DNA controls is done. In lieu of DNA controls, the proposed poisson-gamma (PG) model approach uses an estimate of bias and demonstrates higher sensitivity and lower error rates than previously proposed conventional binomial test. Especially, allelic imbalance is largely related with the perspectives of either cis- and trans-effects from the regulatory sequences in regulatory regions on a gene or coding regions of trans acting factors or through indirect or epistatic effects. Basically, from a common cellular environment two alleles in a diploid individual are expressed. Thus, both alleles from a common cellular environment can vary in regulatory sequence, though, share a common pool of trans factors suggesting that allelic imbalance (AI) between alleles in a common cellular environment deduce functional variability between alleles in cis-regulatory regions.

On the contrary, comparison of the identical allele in different cellular environments, for instance, between genotypes deduces the differences in trans regulation. Previous contributions for detection of AI focused on the limited number of genes and genotypes, moreover, they generally assumed binomial or chi-squared test to test whether allelic specific read counts are violated from the major underlying condition that null expectation is satisfied with there is no AI, namely, two alleles are expressed equally called as the expected proportion. Yet, those naive approaches do not necessarily have the correct error variance. In order to reduce biases by various sources in RNA-seq experimental settings, several Bayesian strategies have been proposed in previous literatures (Aryee *et al.*, 2009; Hardcastle and Kelly, 2010; Leon-Novelo *et al.*, 2014; Lin *et al.*, 2003; Nariai *et al.*, 2013; Shen *et al.*, 2012; Stegle *et al.*, 2010; Vardhanabhuti *et al.*, 2013; Zhao *et al.*, 2008; Zheng and Chen, 2009). Bayesian techniques have been incorporated in the different ways to deal with such biases. Biases are present when aligning to a single reference, a single reference with SNPs masked and multiple references. In order to determine allelic specific read counts in which identical amounts of each allele are present in the sample, DNA sequencing of F1 heterozygotes (DNA controls) is applied. Deviations from 0.5 represents bias in the DNA read counts. In the Bayesian passion gamma model of AI, the parameter can be fixed (q) or random (ϕ) and can be used in conjunction with simulation settings by explicitly considering genome ambiguity and map bias. In the evaluation of this method, the PG model with q=1/2 is preferable to a binomial test and consistently showed a better performance in terms of FDR. To identify allelic imbalance, the flexible Bayesian model is proposed by allowing the presence of DNA controls and by using a fixed or random parameter for the estimate of bias. The proposed PG model

$$y_i|\mu,\alpha,\beta_i,q \sim POI(|\mu\alpha\beta_i q) \text{ and } X_i|\mu,\beta_i,q \sim POI(\mu\beta_i(1-q)), \tag{2.24}$$

where μ is the overall mean, a nuisance parameter, $\beta_i, i = 1,\cdots,I$ represents biological replicate variation effect and q is a constant to play a role in PG model similarly as Negative Binomial (NB) model. In particular, if DNA information is available, random bias parameter ϕ is sampled from the posterior for directly fair comparison, PG model with q=simulation (fixed) and with q=DNA controls (fixed). The parameter of major interest is the treatment effect, α. Let

$$\theta = \frac{\mu\alpha\beta_i}{\mu\beta_i + \mu\alpha\beta_i} = \frac{\alpha}{1+\alpha}, \text{ so when there is no AI}, \alpha = 1. \tag{2.25}$$

And

$$E(\frac{y_i}{x_i+y_i}) = E\left(\frac{E(y_i|x_i+\beta_i)}{x_i+\beta_i}\right) = q \tag{2.26}$$

$$\mu \sim Gamma\left(\alpha_\mu = \frac{1}{2}, \beta_\mu = \frac{1}{2}\right), \beta_1, \cdots, \beta_I \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right),$$
$$\alpha \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right), \eta \sim Gamma(a,b) \tag{2.27}$$

such that $E(\eta) = a/b$. Alternatively used previous models, binomial test and negative binomial with DNA controls, p=bias in DNA are further reviewed in this section. In binomial test, let θ be the unknown proportion of reads from the p paternal allele and let n be the total number of reads aligning to the exon, $H_0 : \theta = \frac{1}{2}, H_\alpha : \theta \neq \frac{1}{2}$. We here reject if $|Z| > 1.96$, where

$$z = \frac{\widehat{\theta} - \frac{1}{2}}{\sqrt{\frac{1}{2}\left(\frac{1}{2}\right)/n}} \tag{2.28}$$

where $\widehat{\theta}$ represents the observed proportion of paternal reads. In negative binomial test, the number of reads is random rather than fixed. For the RNA model, we assume that θ is the parameter for the proportion of reads from the paternal allele, y_i and x_i is the number of RNA reads mapped to the paternal and maternal references, respectively, for the replicate i. Likewise, for DNA model, y_i^* and x_i^* is the number of DNA reads mapped to the paternal and maternal references, respectively, for the replicate i.

The RNA model,

$$x_i|y_i, \theta \sim NB(y_i, \theta) \text{ for } i = 1, 2\cdots, I,\ \theta|p \sim Beta((1-p)t, pt) \tag{2.29}$$

The DNA model,

$$x_i^*|y_i^*, p \sim NB(y_i^*, p) \text{ for } i^* = 1, 2\cdots, I^*,\ p \sim Beta(\nu, \nu) \tag{2.30}$$

where the interpretation of parameterization on NB is such that if $\eta \sim NB(k, \epsilon)$ then $\eta \in \{0, 1, \cdots, \}$ denoting that the number of failures before the first k successes with probability of success equal to ε.

To sum up, this method has improved the performance of identification of allelic specific expression and allelic imbalance through the assessment of AI by accounting for systematic errors which can be identified in simulation PG model outperformed when the bias is known and by easily capturing unidentified variant calls resulting from reduced DNA controls. Additionally, another approach for allelic specific expression proposed by Skelly DA *et al* (Skelly *et al.*, 2011) is reviewed in current article and the method is based on a powerful

and flexible hierarchical Bayesian model to combine information across loci by allowing both global and locust specific inferences about allelic specific expression.

This method is implemented in R environment so that researchers can freely download and apply for this method on their own datasets. In the initial setting, this method is done by the basic binomial test and eventually bayesian hierarchical model is applied suggesting that a superior performance than previous binomial test in terms of true discovery rates. In principle, hierarchical modeling approaches in this method are performed for allelic read counts mapping to reference genome at SNP j in gene i, replicate r as Y_{ijr} in the first stage model. These counts are binomially distributed with parameter N_{ijr} (coverage at the SNP) and p_{ij} is from a gene specific beta distribution with parameters α_i and β_i. The second stage allows a flexible assumption in the aspects of variable p_i across all SNPs within gene i.

In beta-binomial distribution,

$$p_i = \frac{\alpha_i}{\alpha_i + \beta_i},\ e_i = \frac{1}{1 + \alpha_i + \beta_i}, \tag{2.31}$$

samples for these parameters are used by MCMC with 500,000 iterations representing the mean and dispersion parameter of allelic specific expression p_i, respectively when e_i approaches zero, then the counts converge to binomially distribution. A two mixture component prior on p_i and e_i,

$$p_i, e_i | \widehat{a}, \widehat{b}, f, g, h, \pi_0 \sim \begin{cases} Beta(\widehat{a}, \widehat{b}) \times Beta(l, \widehat{d}) \text{ with probabiltiy } \pi_0 \\ Beta(f, g) \times Beta(l, h) \text{ with probabiltiy } 1 - \pi_0 \end{cases} \tag{2.32}$$

where $\widehat{a}, \widehat{b}$ are estimated from genomic DNA data and a measure of the noise in read counts due to technical replicates. Finally, the median values of all posterior samples for these parameters are used by MCMC with 500,000 iterations. Hereby, the unique strength of this method is to infer allelic specific expression (ASE) patterns to vary across SNPs within genes, which can lead to the identification of biologically interesting patterns of ASE that have been little known so far and have critical potential to be investigated with other pervasive post-transcriptional modification in RNA-seq.

3. Concluding Remarks

Until recently, RNA-seq analysis has been popularly performed in a wide range of clinical applications as the cost to sequencing continues to reduce. Compared to microarrays, RNA expression profile has two majorly advantageous features to enable to identify complexity of isoform variability on a single gene to produce various splicing events on transcripts and another nature, allelic specific expression and allelic imbalance resulted from genetic differences in transcriptional rates (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; Leon-Novelo *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; Ng *et al.*, 2014; Nicolae *et al.*, 2011; Niu *et al.*, 2014; Pandey *et al.*, 2013; 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.*, 2010; Wu *et al.*, 2011b; Yalamanchili *et al.*, 2014; Zhang *et al.*, 2014; Zheng and Chen, 2009).

More specifically, the first characteristic is commonly present at higher eukaryotic genomes as demonstrated in the statement of that more than 90 percent of human genes have isoform variants for a single gene. In order to more precisely identify such biological complex phenomena in RNA-seq expression profiles, this article covers a guidance of framework for investigators in the field to discuss more robust Bayesianmethods to increase power of detection and reduce false discovery rates compared to existing methods. Importantly, those intrinsically RNA-seq specific features have been investigated in disease progression for example, tissue (or condition) specifically differential expression whose aberrant patterns are directly and/or indirectly influenced by the causal and consequential outcome for a particular disease of major interest.

To efficiently utilize the wealth of RNA-seq high-throughput data, development of more sophisticated methods to uncover the complexity is continuously necessitated. Furthermore, sufficiently sequencing coverage is also required to detect such characteristics that have not been addressed in the previous technology due to limited resolution. Due to scope and limited pages, we could not include all of methods for alternative splicing events and allelic imbalance in the current review, please take a further look at references if you are interested. Introduced robust Bayesian methods presented a better performance, at least equivalent ability to deeply explore and quantify such RNA-seq specific novel features compared to other alternative competing methods. Therefore, robust methodological strategy will accelerate transcriptome studies by inferring more accurate spliced transcript variability and tissue specific (or condition specific) isoform diversity as well as allelic genetic differences in tissues (Chan *et al.*, 2002; Li and Dickson, 1997; Robakis and Georgakopoulos, 2014). Thus, we here have a greatly important take-home message to be addressed in the upcoming years as the future direction, the focus on gene level analysis might be a partial analytical technique in transcriptional regulatory mechanisms, especially, when both is related with disease progression and therapeutic effects as mentioned in earlier 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; Wang *et al.*, 2014; Yalamanchili *et al.*, 2014). Accordingly, followed by differential gene expression analysis, differential expression at transcripts, exons, and isoforms, and allelic specific expression under the particular tissues (conditions) will be a next stage in the ultra-high-throughput community. Thus, rigorous methodological protocols and appropriate analytical pipeline are the priority to more precisely study and identify the features with the rapid advances of current RNA-seq platform.

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., 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.

Chan, S. L., Pedersen, W. A., Zhu, H. and Mattson, M. P. (2002). Numb modifies neuronal vulnerability to amyloid beta-peptide in an isoform-specific manner by a mechanism involving altered calcium homeostasis: Implications for neuronal death in Alzheimer's disease. *Neuromolecular Medicine*, **1**, 55-67.

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., *et al.* (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.

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., *et al.* (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., *et al.* (2010). Microarray analysis of hippocampal CA1 neurons implicates early endosomal dysfunction during Alzheimer's disease progression. *Biological Psychiatry*, **68**, 885-893.

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.

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.

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*, **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., *et al.* (2014). SigFuge: Single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. *Nucleic Acids Research*, **42**, e113.

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.

Leon-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. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493-500.

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., *et al.* (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.

Li, Y.M. and Dickson, D. W. (1997). Enhanced binding of advanced glycation endproducts (AGE) by the ApoE4 isoform links the mechanism of plaque deposition in Alzheimer's disease. *Neuroscience Letters*, **226**, 155-158.

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**, Article8.

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.

Mezlini, A. M., Smith, E. J., Fiume, M., Buske, O., Savich, G. L., Shah, S., Aparicio, S., Chiang, D. Y., 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.

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**, Suppl 10, S5.

Ng, D. W., Shi, X., Nah, G. and Chen, Z. J. (2014). High-throughput RNA-seq for allelic or locus-specific expression analysis in Arabidopsis-related species, hybrids and allotetraploids. *Methods in Molecular Biology*, **1112**, 33-48.

Nicolae, M., Mangul, S., Mandoiu, II and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, **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.
Robakis, N. K. and Georgakopoulos, A. (2014). Allelic interference: a mechanism for trans-dominant transmission of loss of function in the neurodegeneration of familial Alzheimer's disease. *Neurodegenerative Diseases*, **13**, 126-130.
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. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.
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: Downregulation 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.
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*, **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., 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. (2013). A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-Sample RNA-Seq Data. *Statistics in Biosciences*, **5**, 119-137.
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. (2010). Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *Journal of Bioinformatics and Computational Biology*, **8**, 177-192.
Wang, Y., Lupiani, B., Reddy, S. M., Lamont, S. J. and Zhou, H. (2014). RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in

chickens. *Poultry Science*, **93**, 485-493.
Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R. and Zhang, M. Q. (2011a). SpliceTrap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, **27**, 3010-3016.
Wu, Z., Wang, X. and Zhang, X. (2011b). 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.
Zhang, J., Kuo, C. C. and Chen, L. (2014). WemIQ: An accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics*, `doi:10.1093/bioinformatics/btu757`.
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.
Zheng, S. and Chen, L. (2009). A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Research*, **37**, e75.