Bioinformatics, 32(12), 2016, 1823–1831
doi: 10.1093/bioinformatics/btw060
Advance Access Publication Date: 11 February 2016
Original Paper

OXFORD

Gene expression

An empirical Bayes change-point model for identifying 3′ and 5′ alternative splicing by next-generation RNA sequencing

Jie Zhang and Zhi Wei*

Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on September 28, 2015; revised on September 28, 2015; accepted on January 19, 2016

Abstract

Motivation: Next-generation RNA sequencing (RNA-seq) has been widely used to investigate alternative isoform regulations. Among them, alternative 3′ splice site (SS) and 5′ SS account for more than 30% of all alternative splicing (AS) events in higher eukaryotes. Recent studies have revealed that they play important roles in building complex organisms and have a critical impact on biological functions which could cause disease. Quite a few analytical methods have been developed to facilitate alternative 3′ SS and 5′ SS studies using RNA-seq data. However, these methods have various limitations and their performances may be further improved.
Results: We propose an empirical Bayes change-point model to identify alternative 3′ SS and 5′ SS. Compared with previous methods, our approach has several unique merits. First of all, our model does not rely on annotation information. Instead, it provides for the first time a systematic framework to integrate various information when available, in particular the useful junction read information, in order to obtain better performance. Second, we utilize an empirical Bayes model to efficiently pool information across genes to improve detection efficiency. Third, we provide a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative 3′ SS or 5′ SS happens, and/or where it happens. Simulation studies and real data application have demonstrated that our method is powerful and accurate.
Availability and implementation: The software is implemented in Java and can be freely downloaded from http://ebchangepoint.sourceforge.net/.
Contact: zhiwei@njit.edu

1 Introduction

With continued development and improvement, next-generation RNA sequencing (RNA-seq) has become more and more popular, owing to its affordable cost and advantages over existing technologies. In addition, its superiorities in detecting transcripts in undetermined genomic sequences and very accurate digital resolution make it a more attractive technology which is expected to gradually replace microarray technology (Wang *et al.*, 2009). Many compelling biological problems, once hard to tackle, can be solved by new methods built upon RNA-seq. In particular, alternative splice forms play important roles in building complex organisms from a limited number of genes. They provide a major mechanism for enhancing transcriptome and proteome diversity, and critically regulate various biological functions (Kalsotra and Cooper, 2011; Keren *et al.*, 2010). Using RNA-seq, people observe that more than 90% of human genes undergo alternative splicing, a much higher percentage than anticipated (Blencowe *et al.*, 2009; Wang *et al.*, 2008). Of various alternative splice forms, alternative 3′ SS and 5′ SS are particularly important and constitute more than 30% of all AS events as revealed by RNA-seq (Wang *et al.*, 2008). Several studies have found that alternative 3′ SS and 5′ SS events are relevant to many diseases. By analyzing alternative 3′ SS and 5′ SS events, we can obtain precious diagnostic and prognostic information for therapies (Hammond and Wood, 2011; Singh and Cooper, 2012).

Thanks to high-throughput RNA-seq, genome-wide quantitative studies on AS events become feasible (Pan *et al.*, 2008; Wang *et al.*, 2008). Quite a few computational methods have been developed to detect and identify AS events. These methods can be roughly classified into three categories based on their strategies. The first category, represented by Cufflinks (Trapnell *et al.*, 2010), performs differential splicing detection based on transcript quantification, which is the most challenging. Short reads, sampled from RNA-seq, can be aligned to multiple transcripts due to the similarity and overlaps between alternative transcripts (Huang *et al.*, 2013; Li and Dewey, 2011). It makes the expression estimation of individual transcript an undetermined problem. In addition, various sampling biases, including position-specific biases (Bohnert and Rätsch, 2010; Li *et al.*, 2010; Roberts *et al.*, 2011; Wu *et al.*, 2011b) and sequence-specific biases (Roberts *et al.*, 2011; Turro *et al.*, 2011) in the RNA-seq data, incur daunting difficulties for accurate transcript quantification. Consequently, the efficiency of these methods is diminished by the uncertainty in transcript quantification.

The second category of methods aims to detect differential splicing by testing differential expression of the annotated events obtained from existing splicing databases. Representative examples in this category include ALEXA-seq (Griffith *et al.*, 2010), MISO (Katz *et al.*, 2010), MATS (Shen *et al.*, 2012) and SpliceTrap (Wu *et al.*, 2011a). Among them, MISO employs a statistical model to estimate expression of alternatively spliced exons and isoforms (Katz *et al.*, 2010); MATS leverages a Bayesian statistical framework to flexibly test the hypothesis of differential alternative splicing patterns (Shen *et al.*, 2012). These methods may work well when splicing events are well and accurately annotated. They are not applicable to detect novel AS events not cataloged yet in existing annotation databases.

The third category of methods, including DiffSplice (Hu *et al.*, 2013), DEXSeq (Anders *et al.*, 2012) and FDM (Singh *et al.*, 2011), utilizes splice junction read information to overcome the annotation dependency limitation. The performances of these methods are highly dependent on the number and quality of splice junction reads. Sequencing costs often set limits to sequencing depth and coverage of RNA-seq datasets (Sims *et al.*, 2014) and, consequently, performance of these junction read-based methods. Furthermore, the number and quality of aligned splice junction reads will also rely on sequencing technology as well as read-mapping tools (Engström *et al.*, 2013).

Very recently, Wang *et al.* (2014) propose a change-point model which requires no annotation information. It relies on characterizing the coverage change for detecting alternative polyadenylation (APA). In principle, it can be applied for detecting 3′/5′ AS events. However, compared with APA, a key difference for 3′/5′ alternative splicing is that junction read information can be useful and utilized for locating splice sites. For example, a simple strategy for calling 3′/5′ AS events is to identify locations supported by at least N independent splice junction reads with different alignment start positions (Wang *et al.*, 2009). It is noted that when sequencing depth is not enough, a significant proportion of 3′/5′ AS events may not be covered by junction reads, in particular when using a stringent N threshold for ensuring quality. Due to the junction read coverage limitation, there is room for improvement even when sequencing depth is high. Exon read coverage may be used as clues for 3′/5′ AS events. We expect that using both coverage and junction read information will improve both sensitivity and specificity of AS 3′/5′ calls compared to relaying solely on junction reads or read coverage. It is therefore desirable to develop a method that can systematically integrate both junction read information and coverage information.

From a methodology point of view, Wei's method is a frequestist approach and fails to pool and exploit information across many genes under investigation. In addition, it tests only whether there is a change-point, but not where the change-point is. Thus its change-point location estimation does not guarantee any multiplicity control.

In this paper, we develop an empirical Bayes change-point model for identifying 3′/5′ AS events. Our approach requires no annotation information and is applicable to detect novel aberrant splicing events. Compared with previous methods, our approach has several unique merits. First of all, our model does not rely on annotation information. Instead, it provides for the first time a systematic framework to integrate read coverage information and junction read or annotation information, when available, in order to obtain better performance. Secondly, we utilize an empirical Bayes model to efficiently pool information across genes for improving detection efficiency. Thirdly, we provide a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative splicing happens, and/or where it happens. Simulation studies and applications to real data have demonstrated that our method is powerful and accurate.

2 Methods

2.1 Alternative 3′ SS and 5′ SS selection and change-point problem

We formulate alternative 3′ SS and 5′ SS selection as a change-point problem and illustrate it via toy examples in Figure 1. As shown in Figure 1, different mRNA isoforms, with different expression levels, are generated from a single gene through the alternative selection of 3′ SS or 5′ SS. The common regions (constitutive exons), shared by the two isoforms, are expected to have a higher expression level than the extended regions (spliced regions). As a result, for RNA-seq data, we expect that the common regions will have higher short read densities than the extended regions, and the exons with alternative 3′ SS or 5′ SS will have change-points at their 3′ or 5′ splice sites as illustrated in Figure 1. Thus, we can detect exons with alternative 3′ SS and 5′ SS and their splice sites by detecting the change-points where the short read densities change.

We model exon i with length L_i as a sequence of observations $S_i = (x_{i1}, x_{i2}, \ldots, x_{iL_i})$ ordered in position, where x_{ij} is the number of reads (read-count) whose first base mapped to exon i at position j. Following previous change-point models (Barry and Hartigan, 1993; Denison, 2002; Xuan and Murphy, 2007), we assume the change-points divide the sequence of observations into K unknown homogeneous segments, $\Pi = (\Pi_1, \ldots, \Pi_K)$, such that the data is independent across different segments

$$p(S_{i,1:L_i}|\Pi) = \prod_{k=1}^{K} p(S_{i,\Pi_k}).$$

For our alternative 3′ SS and 5′ SS problem, we further assume $K = 1$ or 2, namely, expecting there is at most one change-point in an exon read-count sequence. We introduce variable $\rho_i \in \{0, 1, 2, \ldots, L_i - 1\}$ for sequence S_i to indicate whether there is a change-point and where the change-point is. Specifically, $\rho_i = 0$ indicates there is no change-point and $\rho_i = \tau (\tau > 0)$ means that there is a change-point at position τ, before which read-counts in $S_{i,1:\tau} = (x_{i1}, \ldots, x_{i\tau})$ follow one homogeneous distribution and after which read-counts in

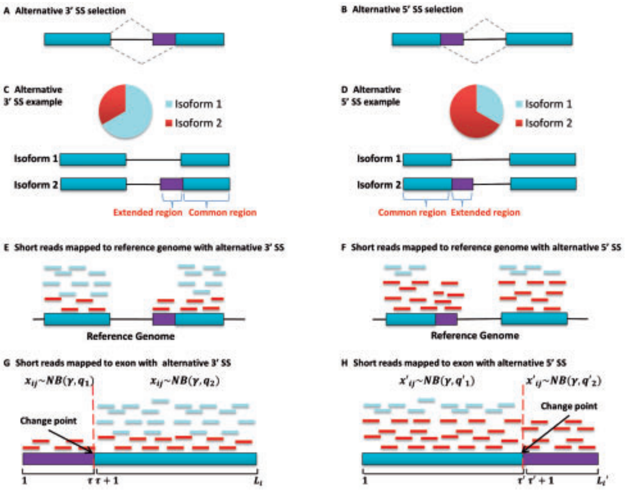
A Alternative 3′ SS selection
B Alternative 5′ SS selection
C Alternative 3′ SS example
Isoform 1
Isoform 2
Isoform 1
Isoform 2
Extended region
Common region
D Alternative 5′ SS example
Isoform 1
Isoform 2
Isoform 1
Isoform 2
Common region
Extended region
E Short reads mapped to reference genome with alternative 3′ SS
Reference Genome
F Short reads mapped to reference genome with alternative 5′ SS
Reference Genome
G Short reads mapped to exon with alternative 3′ SS
x_{ij}~NB(γ, q_1)
x_{ij}~NB(γ, q_2)
Change point
1
τ τ+1
L_i
H Short reads mapped to exon with alternative 5′ SS
x'_{ij}~NB(γ, q'_1)
x'_{ij}~NB(γ, q'_2)
Change point
1
τ′ τ′+1
L_i'

Fig. 1. Illustration and notation of change-point model for alternative 3′ SS and 5′ SS problem. (**A**) and (**B**) show two AS events: alternative 3′ SS and 5′ SS selection, respectively. Blue rectangles represent constitutive exons (common regions) and purple rectangles represent alternatively spliced regions (extended regions). Solid lines and dashed lines indicate the introns and splicing options, respectively. (**C**) and (**D**) are examples of isoforms generated from alternative 3′ SS and 5′ SS selection, respectively. In (C), isoform 1 has a higher expression level, while, in (D), isoform 2 has a higher expression level. (**E**) and (**F**) show the results of mapping short reads to the reference genome, respectively. The reads from isoform 2 are marked as dark red, while reads from isoform 1 are marked as blue. (**G**) and (**H**) show the detailed results of the exons that contain alternative 3′ SS and 5′ SS. Because of the alternative 3′ SS or 5′ SS, the common region shared by the two isoforms has a higher gene expression level than the extended region. Thus, the average number of short reads (read-count) mapped to the common region will be larger than the one for extended region. This generates a change-point at the splice site, which partitions the whole region into two different homogeneous segments with different average read-counts (Color version of this figure is available at *Bioinformatics* online.)

$S_{i,(\tau+1):L_i} = (x_{i(\tau+1)}, \ldots, x_{iL_i})$ follow another homogeneous distribution, as described in details below.

2.2 Negative Binomial-Beta model

Considering the over-dispersion of RNA-seq data, we use a Negative Binomial (NB) distribution to characterize the observed read-counts. As shown in Figure 1, we assume that, when there is no alternative 3′ SS or 5′ SS in exon i, the read-counts across the whole exon i will follow a homogeneous Negative Binomial distribution $\mathrm{NB}(r, q_{i0})$, and when there is a change-point, the alternative 3′ SS or 5′ SS exon contains a splice site at position τ, dividing the read-count sequence into two homogeneous parts governed by two homogeneous Negative Binomial distributions $\mathrm{NB}(r, q_{i1})$ and $\mathrm{NB}(r, q_{i2})$, respectively. Formally, we have

$$x_{ij}|\rho_i \sim \begin{cases} NB(r, q_{i0}), & \text{if } \rho_i = 0 \\ NB(r, q_{i1}), & \text{if } \rho_i = \tau \text{ and } j \le \tau \\ NB(r, q_{i2}), & \text{if } \rho_i = \tau \text{ and } j > \tau \end{cases}$$

Furthermore, we assume q_{ik} follows a Beta distribution $q_{ik} \sim Beta(\alpha, \beta)$, a conjugate prior of the Negative Binomial distribution, in order to characterize fluctuations in technical and biological variation across exons (segments). The hierarchical structure of this Negative Binomial-Beta model is illustrated by Figure 2.

Differently from a conventional Bayesian approach, we will estimate the hyperparameters α and β from the data using an empirical Bayes approach. From the Negative Binomial-Beta model, we have

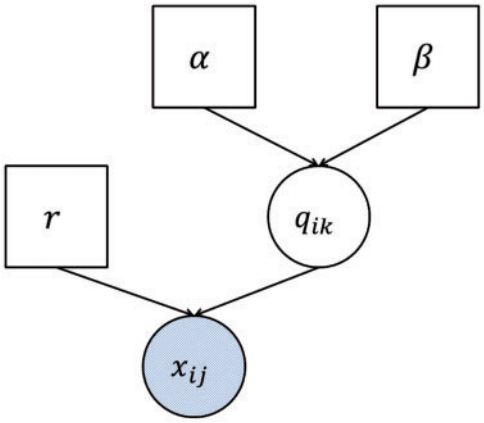

Fig. 2. Hierarchical structure of Negative Binomial-Beta model

$$\begin{aligned} f(x_{ij}|r, q_i) &= \binom{x_{ij}+r-1}{x_{ij}} q_i^r (1-q_i)^{x_{ij}} \\ f(q_i|\alpha, \beta) &= \frac{q_i^{\alpha-1}(1-q_i)^{\beta-1}}{B(\alpha, \beta)}. \end{aligned} \quad (1)$$

Integrating out the unknown exon/segment specific parameter q_i, we derive the likelihood of x_{ij} as

$$\begin{aligned} f(x_{ij}|r, \alpha, \beta) &= \int_q f(x_{ij}|r, q_i) \times f(q_i|\alpha, \beta) \mathrm{d}q \\ &= \binom{x_{ij}+r-1}{x_{ij}} \frac{B(r+\alpha, x_{ij}+\beta)}{B(\alpha, \beta)}. \end{aligned} \quad (2)$$

Since observations in the same segment are independent and identically distributed (i.i.d.), the likelihood for a homogeneous segment $S_{i,j:k} = (x_{ij}, x_{i(j+1)} \ldots, x_{ik})$ can be computed as

$$\begin{aligned} f_0(S_{i,j:k}|r, \alpha, \beta) &= \int_{q_i} \prod_{l=j}^{k} f(x_{il}|r, q_i) \times f(q_i|\alpha, \beta) \mathrm{d}q \\ &= \left[\prod_{l=j}^{k} \binom{x_{il}+r-1}{x_{il}} \right] \\ &\quad \times \frac{B\left((k-j+1)r+\alpha, \sum_{l=j}^{k} x_{il}+\beta\right)}{B(\alpha, \beta)}. \end{aligned} \quad (3)$$

If there is no change-point in the sequence S_i, the likelihood is

$$f(S_i|\rho_i = 0, r, \alpha, \beta) = f_0(S_{i,1:L_i}|r, \alpha, \beta). \quad (4)$$

When there is a change-point at τ, the likelihood is then

$$f(S_i|\rho_i = \tau, r, \alpha, \beta) = f_0(S_{i,1:\tau}|r, \alpha, \beta) \times f_0(S_{i,(\tau+1):L_i}|r, \alpha, \beta). \quad (5)$$

2.3 Prior information and hot points

When no additional information is available, we assume every position has the same prior probability of being a change-point. Suppose that each sequence S_i has a change-point with a prior probability P, and then we assign it equally to each candidate position as

$$Pr(\rho_i; P) = \begin{cases} 1-P & \text{if } \rho_i = 0 \\ \dfrac{P}{L_i - 1} & \text{if } \rho_i = 1, 2, \ldots, L_i - 1. \end{cases}$$

If additional information is available, such as splice junction reads or isoform annotation, we will integrate it and build a more sophisticated model. Specifically, we assign different weights to different candidate positions allowing them to have different prior probabilities as derived from extra information. We define as hot points the positions which are supported by splice junction reads and/or isoform annotation and more likely to be splice sites. The basic strategy is that we assign weight $W \geq 1$, which can be estimated from data or pre-specified by the user, to hot points and weight 1 to ordinary positions. Then the prior probability for each position will be

$$Pr(\rho_i; P, W) = \begin{cases} 1 - P & \text{if } \rho_i = 0 \\ P \times \dfrac{w_{i\rho_i}}{\sum_{j=1}^{L_i-1} w_{ij}} & \text{if } \rho_i = 1, 2, \ldots, L_i - 1, \end{cases}$$

where w_{ij} is the weight assigned to position j in sequence i, (i, j) for short, and

$$w_{ij} = \begin{cases} 1 & \text{if } (i, j) \text{ is ordinary position} \\ W & \text{if } (i, j) \text{ is hotpoint.} \end{cases}$$

This weighting scheme allows flexible weight assigning strategies. It is very useful when the user has different kinds of prior information. Assuming that various information has additive effects on the weight of a candidate position, our model can make full use of all kinds of information. Suppose there are m different kinds of information, we can assign weights to candidate positions as follows

$$w_{ij} = 1 + \beta^{(1)}\delta_{ij}^{(1)} + \beta^{(2)}\delta_{ij}^{(2)} + \cdots + \beta^{(m)}\delta_{ij}^{(m)},$$

where $\beta^{(k)}$ ($k = 1, 2, \ldots, m$) measures the additive effect of information k and $\delta_{ij}^{(k)} = I$ {position (i, j) supported by information k}. Moreover, interactions of different information can be considered and added into the weight assigning procedure by introducing interaction terms $\beta^{(kl)}\delta_{ij}^{(kl)}$ ($k, l = 1, 2, \ldots, m$). For example, given the annotations and splicing reads, a weight assigning strategy can be

$$w_{ij} = 1 + \beta^{(1)}\delta_{ij}^{(1)} + \beta^{(2)}\delta_{ij}^{(2)} + \beta^{(12)}\delta_{ij}^{(12)},$$

where $\beta^{(1)}, \beta^{(2)}$ and $\beta^{(12)}$ measure the additive effects of splice junction reads, isoform annotation and their interactions, respectively.

By assigning different weights to different candidate positions and distinguishing hot points from ordinary ones, we can efficiently leverage domain knowledge to improve our model. Since all parameters ($\beta^{(k)}$) are estimated from data through the empirical Bayes approach, it does not matter if the domain knowledge is dubious or totally wrong. It is noted that there are trade-offs between simple versus sophisticated strategies. Adopting more sophisticated strategy will make better usage of prior information on one hand, but on the other hand, it will introduce more parameters and cause difficulty in parameter estimation. Thus, appropriate strategies need to be chosen to balance these trade-offs for different applications.

2.4 Empirical Bayes estimator

Empirical Bayes estimates combine the Bayesian and frequentist reasoning that the prior probability is estimated frequentistically in order to perform Bayesian inferences (Efron, 2005). This kind of combination not only gives the Bayesian accurate, objective and data-related prior information, but also enables frequentists to obtain more test efficiency in solving scientific problems (Efron, 2010; Zhao *et al.*, 2013). For simplicity and generality, we denote Φ as the set of the parameters for the Negative Binomial-Beta model (r, α and β) and the parameters for characterizing prior information (P and W). The change-point location ρ_i is latent and we sum them out. The maximum likelihood estimation of Φ, applied to the total N sequences, is then

$$\hat{\Phi} = \arg\max_{\Phi} \ \log\left(\prod_{i=1}^{N}\sum_{\rho_i=0}^{L_i-1} Pr(\rho_i|\Phi) f(S_i|\rho_i, \Phi)\right).$$

The optimization algorithm L-BFGS-B (Byrd *et al.*, 1995), a limited-memory modification of the BFGS quasi-Newton method with box constraints, is applied to estimate the parameters Φ. We set appropriate upper and lower bounds for each parameter (e.g. $0 \leq P \leq 1$ for the prior proportion of sequences with change-points). All these procedures are implemented through R (http://www.r-project.org/) and Java.

2.5 Empirical Bayes testing and decision procedure

There are two questions we aim to address:

Q1: Detection, which genes have change-points?
Q2: Identification, where is the change-point, if any?

For Q1, we only care about whether there is a change-point or not and are not concerned about where the change-point locates. For Q2, we aim to find the accurate location of the change-point. In other words, if we correctly detect a sequence with change-point but wrongly locate its position, it is still considered as an error. Given the estimated parameters $\hat{\Phi}$, we have

$$Pr(\rho_i = \tau|S_i; \hat{\Phi}) = \frac{Pr(\rho_i = \tau, S_i; \hat{\Phi})}{\sum_{j=0}^{L_i-1} Pr(\rho_i = j, S_i; \hat{\Phi})}.$$

Let

$$\begin{aligned} \pi_{i0} &= \quad Pr(\rho_i = 0|S_i; \hat{\Phi}) \\ \pi_i^* &= \max\{Pr(\rho_i = \tau|S_i; \hat{\Phi})\} \quad \tau = 1, 2, \ldots, L_i - 1. \end{aligned}$$

For both two questions, we want to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at a nominal level α and find as many sequences with change-points as possible at the same time. To obtain this goal, we propose the following two empirical Bayes testing and decision procedures for Q1 and Q2, respectively.

An empirical Bayes testing and decision procedure for Q1:

1. Order sequences by π_{i0} in an ascending order and denote them by $\pi_0^{(1)}, \pi_0^{(2)}, \ldots, \pi_0^{(N)}$.
2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \pi_0^{(j)} \leq \alpha\}$.
3. Report sequences S_i ($S_i \in \mathcal{G}^{\text{Detection}}$) to have a change-point, where $\mathcal{G}^{\text{Detection}} = \{i : \pi_{i0} \leq \pi_0^{(k)}\}$.

An empirical Bayes testing and decision procedure for Q2:

1. Order sequences by $(1 - \pi_i^*)$ in an ascending order and denote them by $\pi_*^{(1)}, \pi_*^{(2)}, \ldots, \pi_*^{(N)}$.
2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \pi_*^{(j)} \leq \alpha\}$.
3. Report sequences S_i ($S_i \in \mathcal{G}^{\text{Identification}}$) to have a change-point at position τ_i^*, where $Pr(\rho_i = \tau_i^*|S_i; \hat{\Phi}) = \pi_i^*$, and $\mathcal{G}^{\text{Identification}} = \{i : (1 - \pi_i^*) \leq \pi_*^{(k)}\}$.

Through simulation studies, we will show that these two testing procedures are powerful and can control FDR at the nominal level.

3 Results

3.1 Simulation settings

We first perform simulation studies to investigate the numerical performance of the proposed method. We randomly generate $N = 500$ sequences, each with length $L_i = 100$. Then we select $P * N$ sequences to have change-points. We simulate two scenarios, the first one without hot points and the second one with hot points. For the first scenario, we randomly pick one position with equal probability to be the change-point for all the $P * N$ selected sequences. For the second scenario, we set positions 25, 50 and 75 as hot points with weight $W = 32$ while the other points with weight $W = 1$. one half of the selected $P * N$ sequences have change-points at hot points and the other half don't. As a result, we will have $(1 - P) * N + 2 * P * N = (1 + P)N$ homogeneous sequence segments. For these homogeneous sequence segments, we generate q_i for each of them based on a Beta distribution $B(\alpha, \beta)$, which is then used together with r to generate the number of reads aligned to a position based on a Negative Binomial distribution $\mathrm{NB}(r, q_i)$.

We set $r = 10$, $\alpha = \beta = 4$ so that the simulated exons have expression levels close to real data (Mortazavi *et al.*, 2008). We vary P from 0.05 to 0.7 $(P = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7)$ and set the nominal FDR level at 0.1. The simulation was repeated 100 times for each parameter setting, and we report the averaged FDR and sensitivity.

For our empirical Bayes models, we also evaluate two settings. The first one does not consider hot points by setting $W = 1$ rather than estimating it. The second one considers hot points by allowing W to be estimated from data. The method in Wang *et al.* (2014) essentially scans the whole sequence and selects a position exhibiting the most dramatic difference as a potential change-point to be determined by a statistical test. Following their strategy, we implement a frequentist testing procedure as a competing method to be compared with our proposed empirical Bayes method. Specifically, this frequentist method scans the whole sequence and finds a position with the most significant difference as quantified by rank-sum testing statistic. It is an extreme testing statistic and its original P-value may not be valid any more. As a result, this frequentist method could not guarantee multiplicity control. To make comparison, we just rank exons based on their maximum rank-sum testing statistic and then claim the same number of significant exons as our models.

3.2 Simulation results

Figure 3 shows the results for the scenario without hot points. We observe several interesting results. First, our model without considering hot points (NB model without HP), which is ideal for this scenario, can control FDR precisely at the nominal level 0.1 for all settings. Second, our model considering hot points (NB model with HP) can also control the FDR at nominal level 0.1 for all settings. In addition, it demonstrates similar sensitivity as the ideal model which has the information of $W = 1$. This comparable performance suggests that our model considering hot points, which estimates W from data, is more general and robust. Third, both of our empirical Bayes models outperform the frequentist method which demonstrates a higher FDR while with a lower sensitivity. Fourth, the sensitivity for the detection problem (Q1) is much higher than the identification problem (Q2), which is expected, as the latter indeed is more challenging. Figure 4 shows the results for the scenario with hot points simulated. Again, our empirical Bayes models, considering hot points or not, both demonstrate significantly better performance than the frequentist method. Our model considering hot points is the optimal model. It can precisely control FDR at the nominal level and shows the best performance. It is noted that the model without considering hot points erroneously set $W = 1$, and, as a result, it either could not guarantee FDR control for the detection problem (Q1), or has a lower sensitivity than the correct model for the identification problem (Q2).

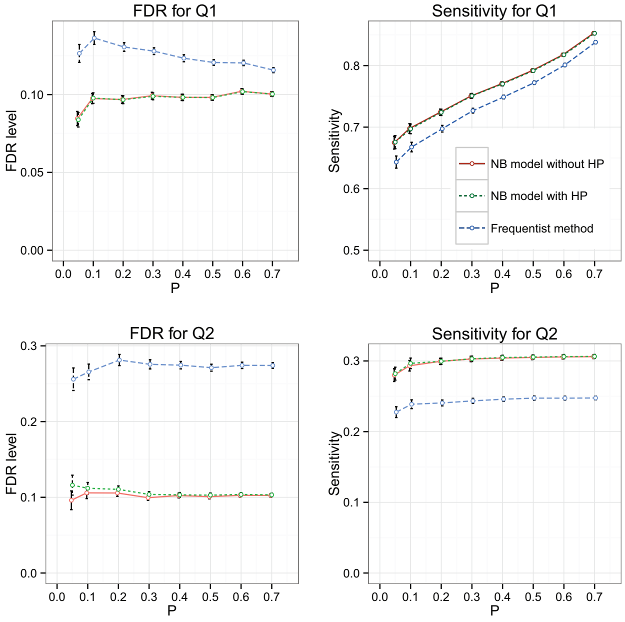

Fig. 3. Results for different methods applied on dataset without hot points. 'NB Model with HP' represents change-point model considering hot points; 'NB Model without HP' represents change-point model without considering hot points

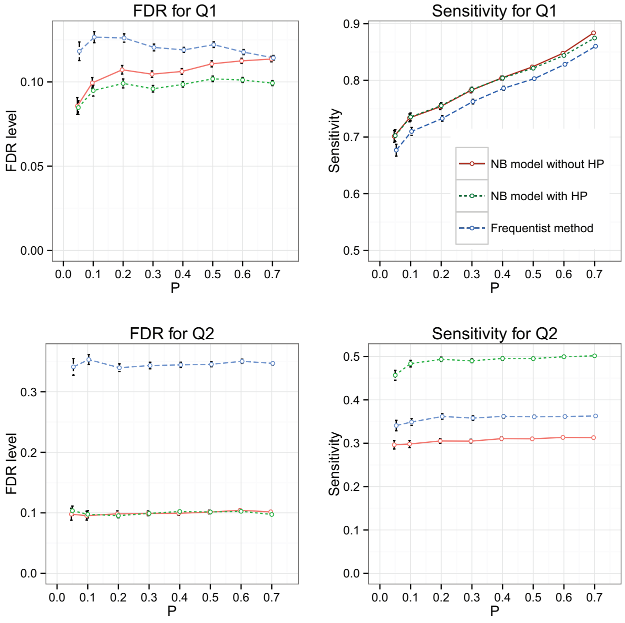

Fig. 4. Results for different methods applied on dataset with hot points. 'NB Model with HP' represents change-point model considering hot points; 'NB Model without HP' represents change-point model without considering hot points (Color version of this figure is available at *Bioinformatics* online.)

From above results, in comparison of the model considering hot points versus the one without considering hot points, we can see that the former is comparable when applied to data without hot points, and better than the latter when applied to data with hot points. Therefore, we conclude that our model considering hot points is robust and superior.

3.3 Integrating junction read information

Splice junction reads are informative for locating 3′/5′ AS events. A simple strategy by counting splice junction reads support has been used for calling 3′/5′ AS events (Wang *et al.*, 2009). It is therefore desirable to exploit this junction read information. Meanwhile, it is noted that additional genes are still being detected even at high read depths of >1 billion reads (SEQC/MAQC-III Consortium, 2014), which suggests that many alternative 3′/5′ AS events may not be covered by junction reads in most current RNA-seq studies and there is room for improvement.

We show that integrating both coverage and junction read information will bring further improvement over the simple strategy that utilizes only junction read information. We use similar hot points simulation settings as in the previous section except that we simulate only one candidate splicing point (hot point) at the middle. We then assume 40, 50, 60 and 70% 3′/5′ splicing events are captured by junction reads (thus the sensitivities achieved by the simple counting strategy), which are used as hot points in our identification model. Table 1 shows the sensitivities of our method averaged over 100 runs. We observe a significant improvement brought by our method in all settings. For example, when junction read coverage is low (40%), our sensitivities are almost doubled. Even when junction read coverage is high (70%), our method can still bring more than 10% improvement.

Table 1. Identification sensitivities using junction reads as hot points

P	Junction coverage			
	40%	50%	60%	70%
0.05	0.710	0.732	0.769	0.795
0.1	0.713	0.750	0.780	0.811
0.2	0.727	0.761	0.795	0.826
0.3	0.730	0.767	0.800	0.834
0.4	0.730	0.767	0.800	0.834
0.5	0.732	0.771	0.805	0.841
0.6	0.740	0.777	0.812	0.850
0.7	0.741	0.780	0.820	0.856

Table 2. Estimated parameters for different samples

Dataset	*P*	*r*	*α*	*β*	*W*
SRR354040	0.354	2.21	1.02	1.22	9.90
SRR354042	0.348	0.86	1.20	1.57	10.00

3.4 Real data application

We then apply our proposed method to analyze a real dataset. Flockhart *et al.* (2012) conducted whole transcriptome RNA-seq to study melanoma cell migration. Their RNA-seq datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) with ID GSE33092. We download their raw reads data from GEO. There are 69 925 376 paired-end reads in the control sample SRR354040 from primary human melanocytes infected with RFP lentivirus, and 62 884 955 paired-end reads in the case sample SRR354042 from primary human melanoma sample (Flockhart *et al.*, 2012). We align the raw reads to the hg19 reference genome using the popular RNA-seq mapping tool Tophat (Trapnell *et al.*, 2009) v1.3.1 with default settings. Exons with short reads found in both samples are used for our analysis. As a result, 62 209 exons from 13 290 distinct genes remain.

We calculate the number of reads starting from each position in every exon and bin them every 5 BPs as one point to reduce the effect of sparsity and noise. In addition, we exploit junction read information in our model by setting junction read-supported positions as hot points, whose weights will be estimated from the data. We apply our model considering hot points on the whole genome. Table 2 shows the estimated parameters for the two samples. We can find that the estimated weights of hot points are bigger than 1 ($\hat{W} > 1$), which indicates that the positions supported by junction reads do have higher prior probabilities than other locations. Our model can capture and make use of this information effectively.

To find biologically meaningful 3′/5′ AS events, we try to detect the exons that have change-points in one sample but not in the other sample. We set the FDR level $\alpha = 0.05$ and detect 7222 such exons. Figure 5 shows four representative examples of the detected exons and their estimated change-point positions. For Figure 5(A) and (B), we observe clear different short read densities before and after the change-points in the melanocyte sample (SRR354040) but not in melanoma sample (SRR354042). In contrast, for Figure 5(C) and (D), there are clear changes in the melanoma sample but not in melanocyte sample. Figure 5(A) shows an example of a detected 5′ AS event which is also covered by junction reads. Figure 5(B) demonstrates that our model can also detect 3′/5′ AS events precisely in the situation when there are no junction reads covering the splice sites. In addition, as shown in Figure 5(C), our model can detect exons with multiple alternative splice sites. The detected AS events in Figure 5(A), (B) and (C) are supported by gene annotations as well, as the detected change-points are close to the splice sites cataloged in UCSC Genes database. We present in Figure 5(D) a potential novel 3′ AS event detected by our model, as it is supported by clear change in read density in the melanoma sample. This potential AS event is not covered by junction reads. The new isoform is not found in RefSeq Genes, UCSC Genes or the AceView database, suggesting that this may be a novel splice site yet to be cataloged.

As a comparison, we also apply the tool developed by (Wang *et al.*, 2014) and the simple strategy, which counts only junction reads, to analyze this dataset. Wei's tool only detects 3366 exons with significant changes between the two samples at the same FDR level, which suggests a lower power compared to our method. For the simple strategy, we call a 3′/5′ AS event if it is supported by one or more junction reads. This strategy detects 796 exons that contain 3′/5′ AS events in one sample but not in the other sample. This improved sensitivity of our method over the simple strategy shows that utilizing junction reads together with read coverage information could obtain a better performance than using junction reads only.

We compare these detected AS events with the isoforms cataloged in the AceView database. When the algorithm detects an AS event and there is one in the matching AceView exon as well, we define this detected AS event as supported by AceView. It is noted that this is not an experimental validation but serves as a proxy to the 'truth'. As summarized in Table 3, 5234 out of the 7222 AS events our method finds are supported by AceView and 1988 AS events are novel, while for Wei's method, 2308 out of the 3366 AS events are supported and 1058 AS events are novel, and for the simple strategy, 772 AS events are supported and 24 AS events are novel. In contrast, for the whole genome, 64.7% exons contain AS events annotated in the AceView database. Here, our model only uses the junction reads

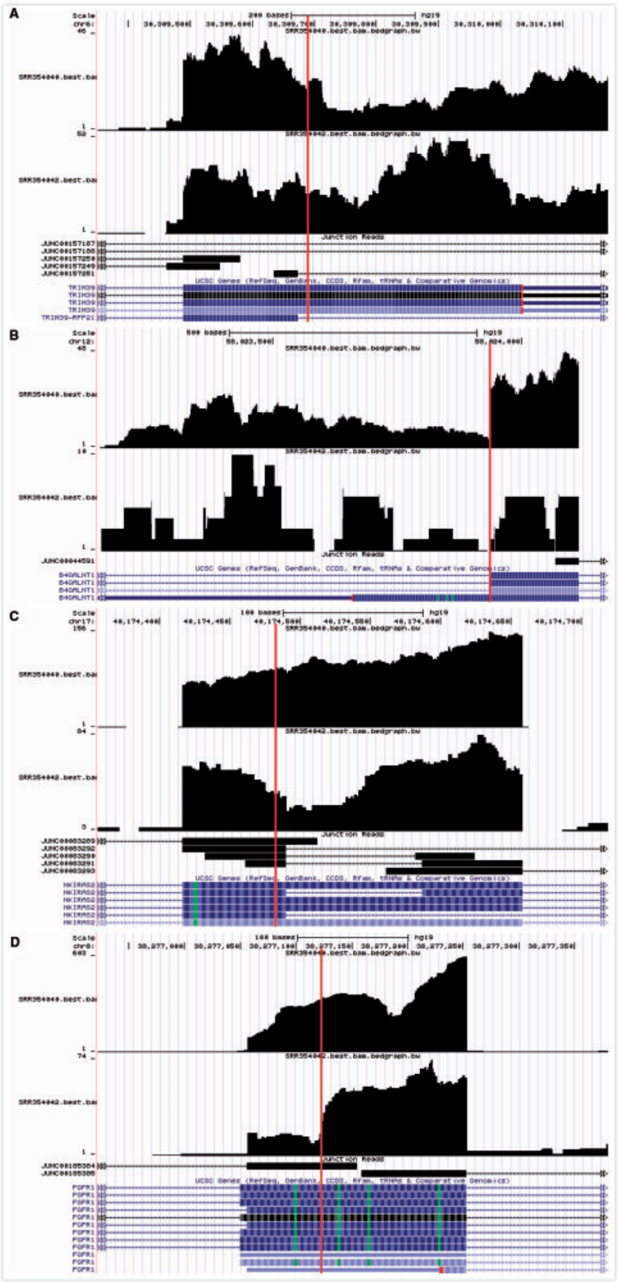
A
B
C
D
Junction Reads
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)

Fig. 5. Examples of detected exons and their splice sites viewed through UCSC Genome Browser together with junction reads and UCSC gene annotations. The red vertical lines indicate the change-points predicted by our model (Color version of this figure is available at *Bioinformatics* online.)

as side information and leaves the AceView annotations for evaluation purpose. The exons reported by our method have a statistically higher supporting rate of 72.5% (*P*-value= 0).

Finally, following Wang *et al.* (2014), we also conduct the gene set enrichment analysis (GSEA) of the genes with 3′/5′ AS events reported by our model in order to evaluate the results from a systems biology point of view. We use the canonical pathways definitions (Version 4) downloaded from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). We identify 12 significantly enriched pathways as shown in Table 4 at an FDR level of 0.05. Interestingly, many of them are relevant to melanoma or cancer. For example, recent studies demonstrate that activation of the Smad2/3 pathway has inhibitory effects on tumor cell plasticity of melanoma (Pardali *et al.*, 2011). Regulation of the actin cytoskeleton contributes to cancer cell migration and invasion (Yamaguchi and Condeelis, 2007). The Notch signaling pathway plays a key role in melanoma growth and progression (Bedogni, 2014). The meaningful GSEA results from a systems biology point of view provide supplementary support to the accuracy of our results. They may provide insight into the role of 3′/5′ AS in these pathways.

Table 3. Results for real data experiment

Method	# Total detected AS	# Novel AS	# AS supported by AceView	Supporting rate
EB Change-point	7222	1988	5234	72.5%
Frequentist Method	3366	1058	2308	68.6%
Simple Strategy	796	24	772	97.0%

Table 4. Gene set enrichment analysis results

Canonical pathway	*P*-value
PID_SMAD2_3NUCLEAR_PATHWAY	3.82E−05
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	7.02E−05
PID_MET_PATHWAY	1.54E−04
KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	1.68E−04
SIG_BCR_SIGNALING_PATHWAY	2.19E−04
REACTOME_DEVELOPMENTAL_BIOLOGY	2.24E−04
BIOCARTA_VDR_PATHWAY	2.30E−04
PID_NECTIN_PATHWAY	2.47E−04
KEGG_PATHWAYS_IN_CANCER	2.50E−04
KEGG_FOCAL_ADHESION	2.94E−04
KEGG_NOTCH_SIGNALING_PATHWAY	4.06E−04
PID_HES_HEY_PATHWAY	4.06E−04

4 Conclusion and discussion

We propose an empirical Bayes change-point model to identify 3′/5′ AS events. Simulation studies and real data application have demonstrated that our method is powerful, accurate and efficient for analyzing the next-generation RNA sequencing data. Compared with previous methods, our approach does not rely on annotation information. Instead, it provides for the first time a systematic framework to characterize coverage change while being capable of integrating other information, in particular the junction read information which is very helpful for detecting 3′/5′ AS events.

We utilize an empirical Bayes model to efficiently pool information across genes. Our Negative Binomial-Beta model, which allows the over-dispersion in the real data, could estimate the hyperparameters from data efficiently. This makes our model more powerful compared with frequentist methods, as our model applies Bayesian inference (Zhao *et al.*, 2013). Since the hyperparameters are estimated frequentistically from data, it also overcomes the defects of subjective priors of Bayesian methods.

We provide a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative splicing happens, and/or where it happens. This gives users more flexibility in solving real problems. When exact splice sites are hard to determine, we could choose to only report the exons that contain alternative splicing. In addition, a Bayesian confidence interval for the splicing point can be constructed based on the posterior probabilities calculated by our method if it is of the user's particular interest.

Recent studies have revealed that RNA-seq data sampled from the transcriptome exhibit various biases, including position-specific and sequence-specific biases (Huang *et al.*, 2013). These biases may incur great difficulties in detecting 3′/5′ AS events and cause false positive reports. When applying our model, the user may circumvent this problem by using additional control samples and doing further filtering. For example, an XOR filtering, $(\rho_i^{\text{case}} > 0)$ XOR $(\rho_i^{\text{control}} > 0)$, or an even sophisticated filtering $|\rho_i^{\text{case}} - \rho_i^{\text{control}}| > \text{a}$ threshold, may be applied.

Our testing and decision procedures allow the user to solve either the detection problem (having or not a change-point) or the identification problem (locating the change-point), but not both simultaneously. The identification problem is more informative as it provides location estimation, but it is also more challenging leading to a lower sensitivity. The detection problem is the converse. To take advantage of both, an idea is to do identification when signal is strong enough and do detection when signal is weak. We are working on this extension within a multiple testing framework and have obtained some promising theoretical results (Sun and Wei, 2015).

It is hard to precisely evaluate the accuracy of the real data analysis results without experimental validation. We use the number of exons detected vs. the number of exons with AceView 'Support' as an implicit indication of the validity of the results. We also conduct GSEA and use the meaningful enriched pathway results as supplementary support to the accuracy of the results. It is noted that both of them are not an experimental validation. We believe that as our tool becomes popular among biologists, a rigorous accuracy validation will be possible and available in the near future.

We assume either one or no alternative 3′/5′ splice sites, represented as the alternative model (M_1) and the null model (M_0), respectively. It is noted that there can be more than one alternative 3′/5′ splice site. In principle, we may extend our model to search for more change-points. To do so, we can replace M_1 (Equation 5) with a more general product partition model M_K allowing $\leq K$ change-points (Barry and Hartigan, 1992). Given multiple change-points, M_1 will fit less well but still better than M_0 assuming no change-points. The null hypothesis (no change-points) we test remains the same. Thus, our testing procedures would remain valid, though not optimal, under some circumstances. Our method may not always be optimal when the signal is subtle. However, we do not expect many such instances. Moreover, if they are interested, the users may manually inspect further those identified genes to see whether there is more than one change-point and where they are. In addition, seeking the optimal model for multiple change-points would impose great computational cost. Computational time is linear to scan for one possible change-point, and becomes factorial when considering multiple change-points. There is also a caveat of over fitting to consider. Because of these implications, the potential gain may not necessarily warrant seeking a perfect model. We would leave this extension for multiple change-points to future work.

Acknowledgement

The authors thank Christopher J. Cardinale for proofreading and all reviewers and the area chair for their constructive comments, which have greatly helped to improve the presentation of the article.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Barry,D. and Hartigan,J.A. (1992) Product partition models for change point problems. *Ann. Stat.*, 260–279. pages

Barry,D. and Hartigan,J.A. (1993) A Bayesian analysis for change point problems. *J. Am. Stat. Assoc.*, **88**, 309–319.

Bedogni,B. (2014) Notch signaling in melanoma: interacting pathways and stromal influences that enhance notch targeting. *Pigm. Cell Melanoma Res.*, **27**, 162–168.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*., **57**, 289–300.

Blencowe,B.J. *et al.* (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.*, **23**, 1379–1386.

Bohnert,R. and Rätsch,G. (2010) rquant. web: a tool for rna-seq-based transcript quantitation. *Nucleic Acids Res.*, **38**, W348–W351.

Byrd,R.H. *et al.* (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Denison,D.G. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, vol. **386**. John Wiley & Sons, Hoboken, New Jersey.

Efron,B. (2005) Bayesians, frequentists, and scientists. *J. Am. Stat. Assoc.*, **100**, 1–5.

Efron,B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. **1**. Cambridge University Press, Cambridge, England.

Engström,P.G. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.

Flockhart,R.J. *et al.* (2012) BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res.*, **22**, 1006–1014.

Griffith,M. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843–847.

Hammond,S.M. and Wood,M.J. (2011) Genetic therapies for RNA mis-splicing diseases. *Trends Genet.*, **27**, 196–205.

Hu,Y. *et al.* (2013) Diffsplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.*, **41**, e39–e39.

Huang,Y. *et al.* (2013) A robust method for transcript quantification with RNA-seq data. *J. Comput. Biol.*, **20**, 167–187.

Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, **12**, 715–729.

Katz,Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.

Keren,H. *et al.* (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.

Li,B. and Dewey,C.N. (2011) Rsem: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.*, **12**, 323

Li,B. *et al.* (2010) RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

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

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Pardali,E. *et al.* (2011) Critical role of endoglin in tumor cell plasticity of ewing sarcoma and melanoma. *Oncogene*, **30**, 334–345.

Roberts,A. *et al.* (2011) Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, R22

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.*, **32**, 903–914.

Shen,S. *et al.* (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.*, **40**, e61

Sims,D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.

Singh,D. *et al.* (2011) Fdm: a graph-based statistical method to detect differential transcription using RNA-seq data. *Bioinformatics*, **27**, 2633–2640.

Singh,R.K. and Cooper,T.A. (2012) Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.*, **18**, 472–482.

Sun,W. and Wei,Z. (2015) Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika*, **102**, 267–280.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Turro,E. *et al.* (2011) Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biol.*, **12**, R13.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,W. *et al.* (2014) A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics*, **30**, 2162–2170.

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

Wu,J. *et al.* (2011a) Splicetrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, **27**, 3010–3016.

Wu,Z. *et al.* (2011b) Using non-uniform read distribution models to improve isoform expression inference in RNA-seq. *Bioinformatics*, **27**, 502–508.

Xuan,X. and Murphy,K. (2007). Modeling changing dependency structure in multivariate time series. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 1055–1062. ACM.

Yamaguchi,H. and Condeelis,J. (2007) Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Res.*, **1773**, 642–652.

Zhao,Z. *et al.* (2013) An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Ann. Appl. Stat.*, **7**, 2229–2248.