Biostatistics (2012), **13**, 4, *pp.* 734–747
doi:10.1093/biostatistics/kxs001
Advance Access publication on February 21, 2012

C

Modeling RNA degradation for RNA-Seq with applications

LIN WAN

Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA and Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

XITING YAN

Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520, USA

TING CHEN, FENGZHU SUN*

Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA and Tsinghua National Laboratory for Information Science and Technology/Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China

fsun@usc.edu

SUMMARY

RNA-Seq is widely used in biological and biomedical studies. Methods for the estimation of the transcript's abundance using RNA-Seq data have been intensively studied, many of which are based on the assumption that the short-reads of RNA-Seq are uniformly distributed along the transcripts. However, the short-reads are found to be nonuniformly distributed along the transcripts, which can greatly reduce the accuracies of these methods based on the uniform assumption. Several methods are developed to adjust the biases induced by this nonuniformity, utilizing the short-read's empirical distribution in transcript. As an alternative, we found that RNA degradation plays a major role in the formation of the short-read's nonuniform distribution and thus developed a new approach that quantifies the short-read's nonuniform distribution by precisely modeling RNA degradation. Our model of RNA degradation fits RNA-Seq data quite well, and based on this model, a new statistical method was further developed to estimate transcript expression level, as well as the RNA degradation rate, for individual genes and their isoforms. We showed that our method can improve the accuracy of transcript isoform expression estimation. The RNA degradation rate of individual transcript we estimated is consistent across samples and/or experiments/platforms. In addition, the RNA degradation rate from our model is independent of the RNA length, consistent with previous studies on RNA decay rate.

Keywords: EM algorithm; Gene expression; Next generation sequencing; RNA degradation; RNA-Seq.

*To whom correspondence should be addressed.

1. Introduction

High throughput transcriptome sequencing (RNA-Seq) is widely used in biological and biomedical studies (Wang *and others*, 2009; Hawkins *and others*, 2010). Compared to microarrays, RNA-Seq has shown superior accuracy in the measurement of gene expression levels (Marioni *and others*, 2008; Mortazavi *and others*, 2008), and it has shown great promise in the study of alternative splicing (Wang *and others*, 2008; Hawkins *and others*, 2010). Computational/statistical methods have been developed for the quantification of transcript abundance using RNA-Seq data to exploit this development (see Pachter, 2011, for a recent review).

Among these methods for the quantification of transcript abundance, many of which are based on the assumption that the short-reads generated by RNA-Seq are uniformly sampled from their transcripts (Jiang and Wong, 2009; Feng *and others*, 2011; Li, Ruotti, *and others*, 2010; Richard *and others*, 2010; Trapnell *and others*, 2010; Katz *and others*, 2010). However, increasing evidences have shown that the short-reads generated by RNA-Seq are nonuniformly sequenced from their transcripts (Wang *and others*, 2009; Pepke *and others*, 2009), and that ignoring such nonuniformity of the short-read distribution in their methods will significantly reduce the accuracy of estimated transcript abundance (Wu *and others*, 2011; Roberts *and others*, 2011). Thus, accurately modeling and efficiently adjusting biases induced by the nonuniformity of short-read distribution are vital for the accurate estimation of transcript abundance.

The nonuniformity of the short-read distribution is attributed to various biases present in RNA-Seq. Previous studies revealed that random primers used in library preparation procedures can induce a bias, which depends on the local sequence content of the transcript (Hansen *and others*, 2010). Such sequence-dependent bias plays a role in the formation of the local nonuniformity of the short-read distribution along transcript and has been modeled and adjusted using the local sequence content of the transcripts (Hansen *and others*, 2010; Li, Jiang, *and others*, 2010; Turro *and others*, 2011; Roberts *and others*, 2011). Another more serious bias is induced by RNA degradation, resulting in a position-dependent pattern. The RNA degradation in RNA-Seq, which is precisely described as the partial RNA degradation and the incomplete extension of RNA during amplification in RNA-Seq (Pepke *and others*, 2009), plays an important role in the formation of the nonuniformity of the short-read distribution in which the degraded and/or the incomplete extended part of the transcript is less likely to be sequenced. Especially in RNA-Seq data which are prepared by complementary DNA (cDNA) fragmentation method, it has been shown that, because of RNA degradation, short-reads tend to be significantly generated more toward the 3' end, followed by an exponential decrease toward the 5' end of the transcript (Wang *and others*, 2009; Pepke *and others*, 2009). RNA degradation is an innate characteristic of RNA molecules, and it is therefore reflected in both RNA-Seq data and the expression microarray data (Archer *and others*, 2006). We showed in this study that the bias induced by RNA degeneration is present in different RNA-Seq platforms.

To adjust the position-dependent bias of the short-read distribution in RNA-Seq, several methods have been developed to incorporate an adjustment based on the empirical distribution of the short-reads along the transcripts (Li, Ruotti, *and others*, 2010; Howard and Heber, 2010; Wu *and others*, 2011; Roberts *and others*, 2011). In brief, a weight can be used to adjust the positional bias for each exon/nucleotide of the transcripts according to the estimated empirical distribution, and the empirical distribution of the short-reads along the transcripts is generally estimated by binning the short-reads on a group of transcripts at the relative positions, implemented by various strategies in different methods (Li, Ruotti, *and others*, 2010; Howard and Heber, 2010; Wu *and others*, 2011; Roberts *and others*, 2011). Although the estimated empirical distribution depicts the decreasing trend of short-reads from the 3' to the 5' end of the transcripts (Wu *and others*, 2011), it is usually difficult for these empirical distribution based methods to accurately characterize the variability of the nonuniformity in individual transcripts, limiting their efficiencies in the adjustment of the position-dependent bias for individual transcripts. As an alternative, quantitative modeling of RNA degradation in RNA-Seq will be of great help to accurately and efficiently adjust the position-dependent bias.

In this study, we report a quantitative model of RNA degradation, with the objective to characterize and adjust the position-dependent bias in RNA-Seq for individual transcripts. We show that our RNA degradation model can quantitatively characterize the effect of RNA degradation on the short-read distribution of individual transcripts and that it fits the RNA-Seq data quite well. Based on this model, we have also developed a new statistical method to estimate the transcript isoform expression levels and RNA degradation rate. We showed that our statistical method is highly accurate in the transcript isoform expression level estimation based on both simulated and real data. The estimated RNA degradation rates of individual transcripts are shown to be consistent across samples and/or experiments/platforms. Meanwhile, we also demonstrated that RNA degradation rate of transcript is independent of RNA length, similar as for RNA decay rates (defined as the inverse of RNA half-life) indicated by previous studies (Wang *and others*, 2002; Yang *and others*, 2003). Furthermore, our model can easily be extended to consider the sequence-dependent bias in RNA-Seq (see supplementary material available at *Biostatistics* online).

2. METHODS

2.1 *Notations*

Suppose that a given gene g has n transcript isoforms with lengths $\{L_{g1}, L_{g2}, \ldots, L_{gn}\}$ and the gene contains a total of x exons. Following Jiang and Wong (2009), when 2 isoforms share part of an exon, we split the exon into several parts and treat each part as a separate exon. The exons are of lengths $l_{g1}, l_{g2}, \ldots, l_{gx}$. We use i $(1 \leqslant i \leqslant n)$ to index the isoform and j $(1 \leqslant j \leqslant x)$ to index the exon of gene g. We define the index matrix $\mathbf{I}_g = (I_{gij})$ of gene g, with $I_{gij} = 1$ when the ith isoform contains the jth exon and $I_{gij} = 0$, otherwise. It is clear that isoform i's length L_{gi} is equal to $\sum_j I_{gij} l_{gj}$. We denote D_{gij} as the distance (in base pair) from the center position of exon j to the 3' end of isoform i (Figure 1) and $d_{gij} \equiv \frac{D_{gij}}{L_{gi}}$ as the normalized distance $(0 \leqslant d_{gij} \leqslant 1)$ from the center position of exon j to the 3' end of isoform i. The expression levels of the isoforms of gene g in an experiment are $\Theta_g = \{\theta_{g1}, \theta_{g2}, \ldots, \theta_{gi}, \ldots, \theta_{gn}\}$. When gene g has a single transcript $(n = 1)$, we will remove the isoform index i from notations L_{gi}, D_{gij}, d_{gij}, and θ_{gi} as L_g, D_{gj}, d_{gj}, and θ_g for simplicity.

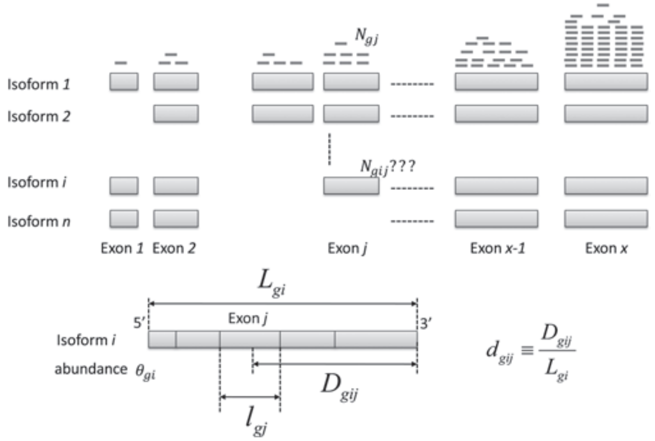

Fig. 1. Notation of the RNA degradation model for genes with multiple isoforms.

2.2 *Modeling of RNA degradation for RNA-Seq*

By mapping the short-reads of an experiment to the reference genome, the numbers of mapped short-reads within each exon $\{N_{g1}, N_{g2}, \dots, N_{gx}\}$ can be obtained. We found that $\{N_{g1}, N_{g2}, \dots, N_{gx}\}$ decreases exponentially from the 3' to the 5' end of the transcript isoform as the result of RNA degradation, especially in RNA-Seq data, which are prepared by cDNA fragmentation method. Accordingly, we developed a mathematical model as follows to quantitatively characterize the effect of RNA degradation on short-read distribution for individual genes and their isoforms.

When gene g has only a single transcript ($n = 1$) and it can be assumed that the gene does not contain overlapping regions with other genes, our RNA degradation model is as follows

$$\frac{N_{gj}}{l_{gj}} = c\theta_g\, \mathrm{e}^{-\alpha_g \frac{D_{gj}}{L_g}} = c\theta_g\, \mathrm{e}^{-\alpha_g d_{gj}},\ \ j = 1, 2, \dots, x, \tag{2.1}$$

where d_{gj} is the normalized distance from the center position of exon j to the 3' end of the transcript (single transcript), θ_g is the expression level of the single transcript of gene g, c is a normalization constant, and α_g ($\geqslant 0$) is a normalized RNA degradation rate. We refer to $l_{gj}\, \mathrm{e}^{-\alpha_g d_{gj}}$ as the effective exon length.

When gene g has multiple isoforms ($n > 1$) and it can be assumed that these isoforms do not contain overlapping regions with other genes, the model is different from (2.1) since N_{gj} is a mixture of short-reads from the isoforms of gene g. We denote N_{gij} ($\geqslant 0$) as the number of mapped short-reads in exon j, which belongs to isoform i, and have

$$\frac{N_{gij}}{l_{gj}} = \begin{cases} c\theta_{gi}\, \mathrm{e}^{-\alpha_g d_{gij}}, & \text{if } I_{gij} = 1; \\ 0, & \text{if } I_{gij} = 0; \end{cases} \tag{2.2}$$

where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, x$, and θ_{gi} is the expression level of isoform i; c is the normalization constant and α_g ($\geqslant 0$) characterizes the rate of RNA degradation of gene g. The N_{gij} is unknown and subjects to the constraint $\sum_i N_{gij} = N_{gj}$. Note that the parameter α_g is gene specific, and the isoform-specific RNA degradation rate α_{gi} should be used when the RNA degradation rates of the isoforms of the gene cannot be considered as the same. See supplementary material available at *Biostatistics* online for the estimation of α_{gi}.

2.3 *Statistical estimation of the RNA degradation rate and expression levels of transcript isoforms*

In order to estimate the RNA degradation rate α_g and the expression levels of isoforms $\Theta_g = \{\theta_{g1}, \theta_{g2}, \dots, \theta_{gn}\}$ based on our model (see 2.2), we consider this as a missing value problem. The observed data are $\{N_{g1}, N_{g2}, \dots, N_{gx}\}$ and the missing data are $\{N_{gij} : i = 1, 2, \dots, n;\ j = 1, 2, \dots, x\}$. The missing data and the observed data form the complete data.

Let $w_{gi} \equiv \sum_j I_{gij} l_{gj}\, \mathrm{e}^{-\alpha_g d_{gij}}$ (the effective length of isoform i of gene g) and $F_g \equiv \sum_{i,j} I_{gij}\theta_{gi} l_{gj}\, \mathrm{e}^{-\alpha_g d_{gij}} = \sum_i \theta_{gi} w_{gi}$. Note that a short-read of gene g has a probability $p_{ij} = I_{gij}\theta_{gi} l_{gj}\, \mathrm{e}^{-\alpha_g d_{gij}} / F_g$ of belonging to isoform i and exon j. Therefore, the likelihood of the complete data for gene g is proportional to

$$L_g = \prod_{i,j} \left(\frac{\theta_{gi} w_{gi} \frac{l_{gj}\, \mathrm{e}^{-\alpha_g d_{gij}}}{w_{gi}}}{\sum_h \theta_{gh} w_{gh}} \right)^{I_{gij} N_{gij}}, \tag{2.3}$$

and the log-likelihood function is

$$\log L_g = \sum_{i,j} I_{gij} N_{gij} \left(\log(\theta_{gi} w_{gi}) - \log\left(\sum_h \theta_{gh} w_{gh} \right) \right) + \sum_{i,j} I_{gij} N_{gij} \log\left(\frac{l_{gj}\, \mathrm{e}^{-\alpha_g d_{gij}}}{w_{gi}} \right). \tag{2.4}$$

Note that we want to maximize $\log L_g$ as a function of $(\theta_{g1}, \theta_{g2}, \ldots, \theta_{gn}, \alpha_g)$ with the constraint $\sum_{i=1}^{n} \theta_{gi} = 1$.

We developed an expectation–maximization (EM) algorithm to maximize this log-likelihood function $\log L_g$ as follows. In our implementations, we set the initial values as $\alpha_g^{(0)} = 0$ and $\theta_{gi}^{(0)} = 1/n$. Given the current values of $\Theta_g^{(t)} = (\theta_{g1}^{(t)}, \theta_{g2}^{(t)}, \ldots, \theta_{gx}^{(t)})$ and $\alpha_g^{(t)}$ at the tth step, we take the expected value of $\log L_g$ in the E-step and have

$$N_{gij}^{(t)} = E(N_{gij} | (N_{g1}, N_{g2}, \ldots, N_{gx}), \Theta_g^{(t)}, \alpha_g^{(t)}) = \frac{N_{gj} I_{gij} \theta_{gi}^{(t)} \,\mathrm{e}^{-\alpha_g^{(t)} d_{gij}}}{\sum_h I_{ghj} \theta_{gh}^{(t)} \,\mathrm{e}^{-\alpha_g^{(t)} d_{ghj}}}, \tag{2.5}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, x$.

In the M-step of the EM algorithm, we maximize (2.4) with respect to Θ_g and α_g by replacing N_{gij} with $N_{gij}^{(t)}$.

If we let $\beta_{gi} \equiv \theta_{gi} w_{gi} / \sum_h \theta_{gh} w_{gh}$, (2.4) will be reparametrized as a function of $(\beta_{g1}, \ldots, \beta_{gn}, \alpha_g)$, where the first term of the right-hand side does not contain α_g and the second term does not contain $(\beta_{g1}, \ldots, \beta_{gn})$. Therefore, to estimated α_g, we only need to maximize

$$G_t(\alpha_g) = \sum_{i,j} I_{gij} N_{gij}^{(t)} \log\left(\frac{l_{gj} \,\mathrm{e}^{-\alpha_g d_{gij}}}{w_{gi}} \right),$$

with respect to α_g (note that w_{gi} is a function of α_g). We apply the Newton–Raphson method to maximize $G_t(\alpha_g)$ and thus estimate $\hat{\alpha_g}^{(t+1)}$ (see supplementary material available at *Biostatistics* online for details). The $\hat{w}_{gi}^{(t+1)}$ is determined when $\hat{\alpha}_g^{(t+1)}$ is known.

To estimate Θ_g, we then maximize the first term of the right-hand side of (2.4) and have

$$\hat{\beta}_{gi}^{(t+1)} = \frac{\sum_j I_{gij} N_{gij}^{(t)}}{N_g}, \quad i = 1, 2, \ldots, n,$$

where N_g is the total number of mapped short-reads from gene g and $N_g = N_{g1} + N_{g2} + \cdots + N_{gx}$. To make the model identifiable, we apply the constraint $\sum_i \theta_{gi} = 1$ as in Trapnell *and others* (2010) such that

$$\hat{\theta}_{gi}^{(t+1)} = \frac{\hat{\beta}_{gi}^{(t+1)} \frac{1}{\hat{w}_{gi}^{(t+1)}}}{\sum_h \hat{\beta}_{gh}^{(t+1)} \frac{1}{\hat{w}_{gh}^{(t+1)}}}, \quad i = 1, 2, \ldots, n. \tag{2.6}$$

Refer to Lemma 14 of Trapnell *and others* (2010) for the derivations. We refer to the above RNA degradation-based method as RD.

2.4 *Data and data processing*

We demonstrate our model and method using 2 RNA-Seq data sets from 2 independent laboratories with 2 different sequencing platforms. Data set I is from the Human Body Map 2.0 Project sequenced by Illumina. This data set was generated by the Illumina HiSeq 2000 platform and contains RNA-Seq data of 16 different human tissues. For each tissue sample, the sequence library was prepared by using the standard poly(A)-selected messenger RNA (mRNA). The project sequenced the mRNA of each tissue by one lane with single-end 75-bp sequencing reads. Data set II is from Marioni *and others* (2008), containing

RNA-Seq data of human liver and kidney tissues. It was sequenced by the Illumina GA platform with a sequencing read length of 36 bp. Data set II was generated using the standard poly(A)-selected mRNA library. It was downloaded from Sequence Read Archive (http://trace.ncbi.nlm.nih.gov/Traces/home/) with accession number SRX000571 (liver).

Using Bowtie (version 0.12.5) (Langmead *and others*, 2009), we mapped the 75 and 36 bp short-reads to the human genome (hg18; http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg18), allowing, at most 3 (for 75 bp reads) and 2 (for 36 bp reads) mismatches. The RefSeq annotation (Pruitt *and others*, 2009) was used as the annotations of genes and their isoforms (downloaded from University of California Santa Cruz Genome Browser; Kent *and others*, 2002; on November 7, 2010). We only kept the uniquely mapped short-reads and did not consider the junction reads as in Wu *and others* (2011) for simplicity. We checked the case when junction reads are included to see whether our RNA degradation model can still fit the data in supplementary material available at *Biostatistics* online.

2.5 *Simulations*

Simulations were implemented to evaluate the accuracy of our method on the estimation of the RNA degradation rate α_g and transcript isoform expression levels. We used the real isoform structures based on the isoform annotations in RefSeq. All the genes with 2–9 isoforms were used. Genes with the same number of isoforms were grouped into the same gene set.

For a given gene g with its annotation, we first randomly generate the relative isoform expression levels θ_{gi} satisfying $0 \leqslant \theta_{gi} \leqslant 1$ and $\sum_{i=1}^{n} \theta_{gi} = 1$. To consider the biological reality that many isoforms of a gene are not expressed, we randomly selected m of the n isoforms as expressed and simulated the expression levels θ_{gi} of the expressed isoforms by generating m random numbers in the unit interval (0,1] and then dividing the m random numbers by their summation; the expression levels θ_{gi} of the remaining isoforms were set as 0.

The negative binomial (NB) model is widely used in the modeling of the RNA-Seq read count data to account for the high variability of the read count data (Oshlack *and others*, 2010). We thus simulated the short-reads count data of exon j belonging to isoform i from the NB distribution as follows

$$N_{gij} \sim \begin{cases} \mathrm{NB}(\mu = c l_{gj} \theta_{gi}\, \mathrm{e}^{-\alpha_g d_{gij}}, \phi), & \text{if } I_{gij} = 1; \\ 0, & \text{if } I_{gij} = 0; \end{cases} \tag{2.7}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, x$, where μ is the mean and ϕ is the dispersion parameter of the NB distribution, and the variance of the NB distribution equals $\mu + \mu^2/\phi$. The c is a normalization constant of the experiment, which reflects the genomic coverage of the experiment. Finally, we sum short-read counts from each isoform and obtain the short-reads count for each exon j as $N_{gj} = \sum_{i=1}^{n} N_{gij}$. The simulated values of N_{gj} are the inputs to our statistical method.

In addition, we also simulated the read counts data by an alternative model different from our RD model, to check the robustness of our RD method in the estimation of isoform expression level and the RNA degradation rate (see supplementary material available at *Biostatistics* online).

3. Results

3.1 *RNA degradation in RNA-Seq and the model*

We developed a mathematical model to quantify the decreasing trend of short-read counts in exons from the 3' to the 5' end of the transcripts, as governed by RNA degradation (Section 2.2). To show that our RNA degradation model for RNA-Seq can fit the RNA-Seq data, we demonstrate using genes, which have a single transcript and determine if they fit (2.1). Since the model for genes with multiple isoforms (2.2) is a general extension of the single-transcript case, the validation of (2.1) can also validate (2.2).

We took the logarithms on both sides of (2.1) (suppose $N_{gj} > 0$) and added an error term such that

$$\log(N_{gj}/l_{gj}) = \log c + \log \theta_g - \alpha_g d_{gj} + \epsilon, \ j = 1, 2, \ldots, x. \tag{3.1}$$

We assumed that the error term ϵ follows a normal distribution with mean zero and standard deviation σ. If the real data fit the model, we see that $\log(N_{gj}/l_{gj})$ and d_{gj} will follow the linear model (3.1) with a slope of "$-\alpha_g$" and an intercept of "$\log c + \log \theta_g$." To confirm this, we selected the human liver tissue sample in Data set I as a demonstration (similar results on Data set II were in supplementary material available at *Biostatistics* online). We first mapped the short-reads of the liver sample to the human genome (hg18). Among the genes with single transcript based on the RefSeq annotation, we chose those genes having a total number of >1000 mapped short-reads in the liver sample. In addition, for each selected gene, we filtered out the exons having no short-reads mapped or having an exon length less than 150 bp. After this filtering, a total of 1882 genes with 3 or more retained exons were kept for our analysis.

For each of the 1882 genes, we calculated $\log(N_{gj}/l_{gj})$ for its exons (retained after the filtering) and then performed a linear regression on $\log(N_{gj}/l_{gj})$ with respect to d_{gj}. Figure 2(a) shows an example of the housekeeping gene ACTB. The ACTB gene has a total of 76 248 mapped short-reads. After our filtering, 4 exons are left with 75 553 mapped short-reads in them. For each of the 4 exons, it is clear that the $\log(N_{gj}/l_{gj})$ decreases with the increase of d_{gj}, which fits a linear model quite well with the $R^2 = 0.9985$ (Figure 2(a)). Similar plots of all 1882 genes are shown in supplementary material available at *Biostatistics* online. Figure 2(b) summarizes the histogram of the R^2 of all regressions on the 1882 genes. Among the 1882 genes, 17.5% have an $R^2 \geqslant 0.9$; 34.5% have an $R^2 \geqslant 0.8$; 80.0% have an $R^2 \geqslant 0.5$; the median of the R^2s of all 1882 genes is 0.7145. Such results suggest that RNA-Seq data can fit our RNA degradation model well.

We showed that when taking the junction reads into account, our RNA degradation model still fits the data well and achieved almost the same results as the case without including the junctions (see supplementary material available at *Biostatistics* online).

3.2 *RNA degradation rate α_g*

Precise measurement of the RNA degradation rate of each gene is useful for quantitative studies of the RNA degradation mechanism. Specifically, our RD model can be used to accurately estimate the RNA degradation rate α_g for individual genes. For genes with a single transcript, the linear model of (3.1) shows that the RNA degradation rate α_g is positive, indicating that the number of short-reads decreases from the 3' to the 5' end of the transcripts. To make sure that the α_gs used here are accurate, we chose 951 genes with $R^2 \geqslant 0.7$ from the 1882 genes we used in the above section. Among the 951 genes, 945 have positive α_g, as expected. The histogram of the α_gs for the 945 genes is shown in Figure 2(c). On the other hand, 6 genes (AGPAT6 (NM_178819), CISD2 (NM_001008388), HMGB3 (NM_005342), MFF (NM_020194), PMPCB (NM_004279), and TMEM195 (NM_001004320)) of the 951 genes have negative α_gs, showing a reverse trend of the RNA degradation where the short-reads decrease exponentially from the 5' to the 3' end of the transcript. We explored genomic regions around the 6 genes to find out the reason. Except for gene AGPAT6 (NM_178819), we found that the other 5 genes have overlapping regions with other genes, a case which violates the assumption of our model.

To check whether the RNA degradation rates α_g of individual genes are consistent across samples and/or experiments/platforms, we compared the estimated values of α_g between (1) the liver and kidney samples in Data set I and (2) liver samples in Data sets I and II. Following similar procedures as above for the liver sample of Data set I, a total of 1269 genes (with single transcript) in the kidney sample of Data set

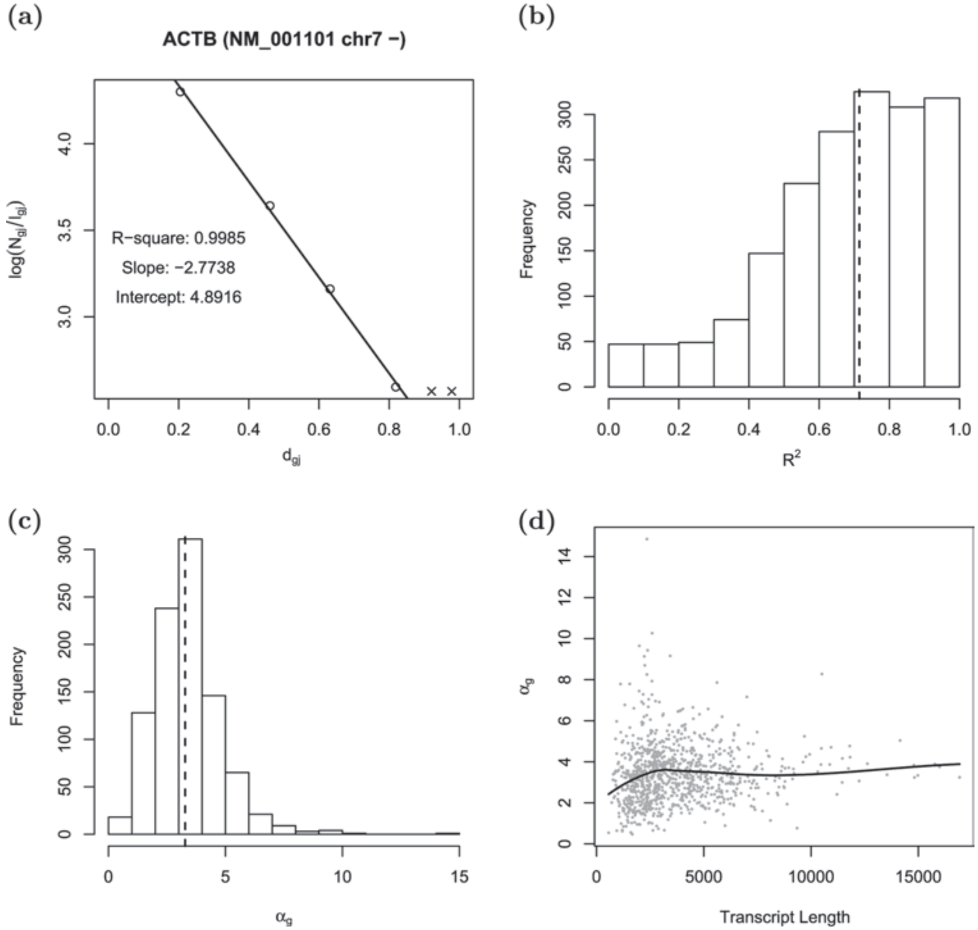

Fig. 2. The RNA degradation model for RNA-Seq. (a) RNA degradation for the gene ACTB. The number of mapped short-reads for each exon divided by exon length decreases exponentially, as the distance of the exon to the 3’ end of the transcript increases; the circles represent the exons, and the solid line is the linear regression result. The crosses show the location of the exons we filtered out. (b) Histogram of the R^2 of the linear regressions on the 1820 genes. The dashed line shows the median of the R^2s. (c) Histogram of the estimated α_gs of the 945 genes with positive α_g and $R^2 \geqslant 0.7$. The dashed line shows the median of the estimated α_gs. (d) The relationship between the value of α_g and transcript length. The gray circles are from 945 genes as in (c); the curve is estimated based on local regression by the loess method on the 945 genes. The loess regression was performed by the R function “loess” with the default setting. All plots and results are based on Data set I.

I are selected, having >1000 mapped short-reads with a positive α_g estimated and $R^2 \geqslant 0.7$. There are 625 common genes in the selected genes from the liver and kidney samples of Data set I. The estimated values of α_g of the 625 common genes based on the 2 samples are highly consistent with a Pearson correlation coefficient $\rho = 0.75$ (Figure 3(a)). We applied similar procedures to select genes from liver sample of Data set II (see supplementary material available at *Biostatistics* online), and the estimated values of α_g of the 603 common genes based on the liver samples from Data sets I and II are still consistent with a Pearson correlation coefficient $\rho = 0.63$ (Figure 3(b)). Such results suggest that the RNA degradation rate α_g is more consistent between samples from within the same experiment and platform than across experiments/isoforms.

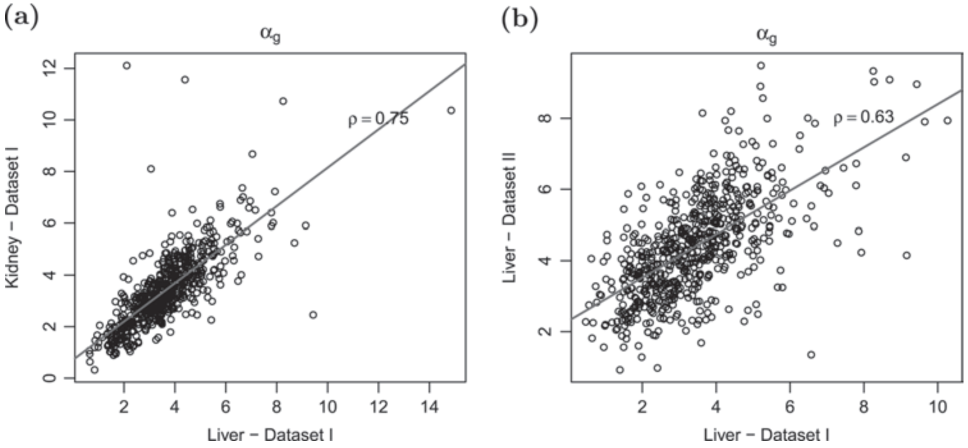

Fig. 3. The consistency of RNA degradation rate. (a) The estimated values of α_g of the 625 common genes based on the liver and kidney samples of Data set I. (b) The estimated values of α_g of the 603 common genes based on the liver samples from Data sets I and II. The solid lines are the linear regression lines between the 2 samples in each plot. The ρ stands for Pearson correlation coefficient.

We also studied the relationship between the RNA degradation rates α_g and transcript lengths using the 945 genes from the liver sample of Data set I (Figure 2(d), the gray circle dots) and found no significant correlation between RNA degradation rate and the transcripts length (Pearson correlation coefficient $\rho = 0.08$ and $R^2 = 0.006$). The local regression showed that RNA degradation rate and the transcript lengths are independent (Figure 2(d)). Furthermore, α_g has great variability at a fixed transcript length (note that when the transcript length >10 000 bp, the samples size is too small to see the variation). Similar results were also obtained from Data set II (see supplementary material available at *Biostatistics* online).

3.3 *Applications to accurately estimate transcript isoform expression levels*

We developed a statistical method based on our RNA degradation model to estimate the isoform expression levels, as well as the RNA degradation rate (Section 2.3). To compare our RNA degradation-based method (RD) with the uniform assumption-based methods (UN), we fixed the $\alpha_g = 0$ during the estimation to make RD degenerate into UN. We did not compare with the empirical distribution-based method by Wu *and others* (2011) because their program has not been publicly available yet. However, we have already shown that the RNA degradation rate α_g can vary greatly, leading to significant variability in the nonuniformity of short-read distribution.

We evaluated our RD method with the UN method on simulated data because we currently lack benchmark RNA-Seq data sets with experimental validated isoform expression levels. The details of our simulation are described in Section 2.5. To ensure that our simulation can approximate the real data, we simulated the test data by (1) using real isoform structures based on RefSeq annotation, (2) generating the relative expression levels θ_{gi} of isoforms such that only 1 or 2 isoforms are expressed within each gene g ($m = 1, 2$) to consider the fact that many isoforms of a gene are not expressed, (3) generating the short-read counts of the exons based on the NB distribution (the mean μ was chosen based on our RD model and the dispersion parameter ϕ was chosen from 1 to 10 with step 1); and (4) choosing parameters α_g based on their estimated values in the RNA-Seq data (according to Figure 2(c), we selected $\alpha_g = 1, 3, 5, 7, 9$). Three normalization constants ($c = 1, 5, 10$) were chosen to reflect different sequence depths.

Following Wu *and others* (2011), we used 2 measurements, the major isoform recovery rate (MIRR) and difference score (DS), to evaluate the accuracy of our methods. The term "major isoform" refers to the isoform with highest expression level among alternatives in a given gene. The MIRR is defined as the percentage of genes whose major isoforms are correctly identified (Wu *and others*, 2011). As MIRR percentage increases, the accuracy of estimation also increases. The DS of gene g is defined as

$$\mathrm{DS} = \sum_{i=1}^{n} \left| \theta_{gi} - \hat{\theta}_{gi} \right|, \tag{3.2}$$

with the range $0 \leqslant \mathrm{DS} \leqslant 2$ (note that the expression levels are relative expression levels of isoform satisfying that $\sum_{i=1}^{n} \theta_{gi} = 1$). As DS decreases, the accuracy of estimation increases.

For different combinations of the parameters α_g and ϕ (the variance of NB model decreases with the increasing of ϕ), we conducted extensive simulations on all genes with 2–9 isoforms and calculated both MIRR and averaged DS for each set of genes with the same number of isoforms. The overall results show that RD significantly outperforms UN by both MIRR and averaged DS (see Figures 4 and 5 for gene

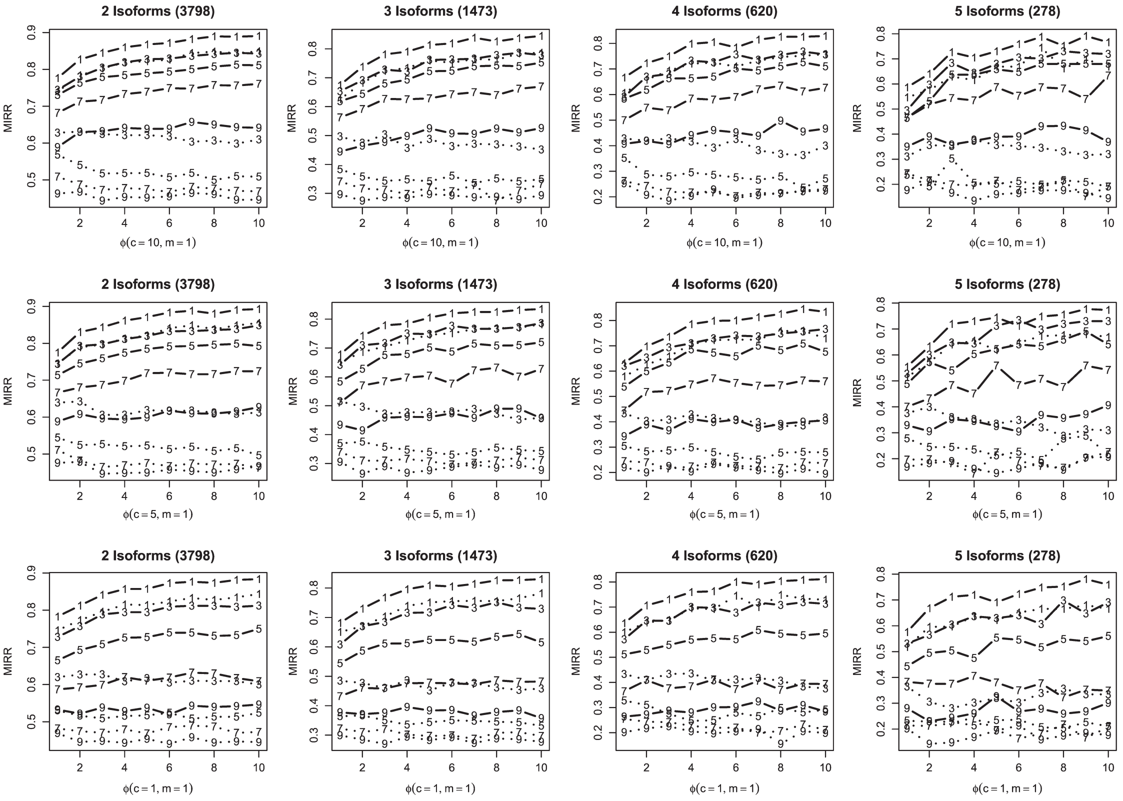

Fig. 4. Simulation results of MIRR under the situation that only one isoform is expressed within each gene ($m = 1$). Each plot shows the MIRRs on the genes with n isoforms ($n = 2, 3, 4, 5$); the gene number is shown in the parentheses of the title in each plot. In each plot, we compare our RNA degradation-based method (RD, shown in solid lines) with the uniform assumption-based method (UN, shown in dotted lines) by testing the simulated data with different combinations of parameters: $\alpha_g = 1, 3, 5, 7, 9$ (shown with dots from "1" to "9"), $\phi = 1, 2, \ldots, 10$, and $c = 1, 5, 10$ (row).

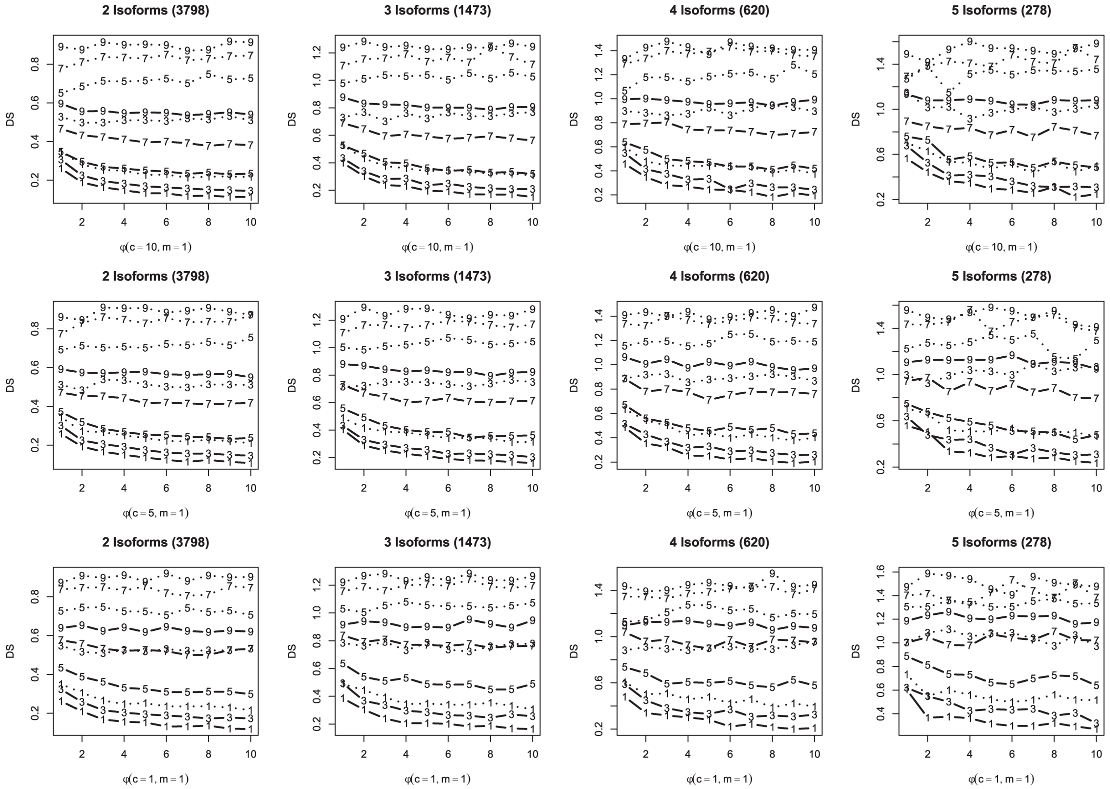

Fig. 5. Simulation results of DS under the situation that only one isoform is expressed within each gene ($m = 1$). Each plot shows the averaged DSs on the genes with n isoforms ($n = 2, 3, 4, 5$); the gene number is shown in the parentheses of the title in each plot. In each plot, we compare our RNA degradation-based method (RD, shown in solid lines) with the uniform assumption-based method (UN, shown in dotted lines) by testing the simulated data with different combinations of parameters: $\alpha_g = 1, 3, 5, 7, 9$ (shown with dots from "1" to "9"), $\phi = 1, 2, \ldots, 10$, and $c = 1, 5, 10$ (row).

sets with 2–5 isoforms and Supplementary Figures 4 and 5 of the supplementary material available at *Biostatistics* online for gene sets with 6–9 isoforms for the case that $m = 1$; see Supplementary Figures 6–9 of the supplementary material available at *Biostatistics* online for the case that $m = 2$). The accuracies of both RD and UN decrease with the increase of isoform number n of each gene. The accuracies of RD increase as the ϕ increase (the variance of the NB model will decrease). The accuracies of RD are not sensitive to the c except that when $\alpha_g \geqslant 7$, the MIRR of RD decreases and the DS of RD increases slightly when c decreases from 10 to 1. When $c < 1$, the coverage for the gene is too low to be considered by our model in real applications. Meanwhile, the estimation of α_g by our method is unbiased when $\phi \geqslant 2$ (see Supplementary Figures 10–17 of the supplementary material available at *Biostatistics* online for some examples).

We checked the performance of our RD method at the situation that RNA degradation follows a different model from our RD model by simulation and found that the RD method still outperform the UN method for most cases in the estimation of isoform expression level (see supplementary material available at *Biostatistics* online). An application of our method to a real example is also described in the supplementary material available at *Biostatistics* online.

4. Discussion

In this study, we developed a mathematical model for the RNA degradation present in RNA-Seq. The model fits the RNA-Seq data quite well for most cases. As we already mentioned in Section 1, the RNA degradation in RNA-Seq is the combined effects of both the innate degradation of RNA molecular and the incomplete extension of RNA molecular during PCR amplification. Pickrell *and others* (2010) have indicated that the sequence contents (e.g. GC contents, which are defined as the proportions of G or C nucleotide of the gene) are related to the PCR amplification efficiency and used the GC contents to adjust the bias of PCR amplification. An RNA degradation model at the molecular level with considering sequence contents is needed for our further understanding of RNA degradation present in RNA-Seq.

Meanwhile, we showed that RNA degradation rate α_g is independent of RNA length. It is important to notice that the RNA decay rate, which is defined as the inverse of RNA half-life, is also independent of RNA length in yeast and human (Wang *and others*, 2002; Yang *and others*, 2003). The RNA degradation rate α_g defined in this study is proportional to the RNA decay rate in Wang *and others* (2002) and Yang *and others* (2003). The molecular mechanisms behind RNA degradation are complicated and still unclear, which is an interesting topic for future research.

Although our RNA degradation model and the estimation method based on it are successful in these applications, both model and method still have limitations. Specifically, we noted that a small portion of genes do not fit our model well (e.g. 20% genes have $R^2 < 0.5$ in the linear regression of (3.1)). Fundamentally, real data are often distorted by various biases and effects during experimentation and data processing, which we have not considered in the current implementation scheme. Several reasons may account for the lack of fit of the model for these genes. First, the method depends on existing isoform annotations, which may not be accurate or complete. Second, our current model requires that the gene not overlap with other genes, which may not be correct for some genes. For example, about 20% human transcripts form sense–antisense gene pairs (Chen *and others*, 2004). We showed that among the 6 genes with estimated negative α_g, 5 of them were found to overlap with other genes. To incorporate these complexities into our model is a topic for future research.

In this study, we only consider RNA-Seq data of single-end short-reads. The paired-end RNA-Seq is also widely used, having advantages in the detection of alternative splicing events compared with the single-end RNA-Seq (Trapnell *and others*, 2010; Katz *and others*, 2010; Salzman *and others*, 2011). RNA degradation still exists in the paired-end RNA-Seq data and may therefore also reduce the accuracy of the estimations of transcript expression levels based on paired-end RNA-Seq. We will therefore extend our current model to accommodate paired-end RNA-Seq in the future.

Software

Software is available on http://www-rcf.usc.edu/ fsun/programs.html.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

Funding

This research was supported by National Institutes of Health (P50 HG 002790 and 1 U01 HL108634). F.S. is also supported by National Natural Science Foundation of China (60928007 and 60805010) and Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross-discipline Foundation.

REFERENCES

ARCHER, K. J., DUMUR, C. I., JOEL, S. E. AND RAMAKRISHNAN, V. (2006). Assessing quality of hybridized RNA in Affymetrix Genechip experiments using mixed-effects models. *Biostatistics* **7**, 198–212.

CHEN, J., SUN, M., KENT, W. J., HUANG, X., XIE, H., WANG, W., ZHOU, G., SHI, R. Z. AND ROWLEY, J. D. (2004). Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Research* **32**, 4812–4820.

FENG, J., LI, W. AND JIANG, T. (2011). Inference of isoforms from short sequence reads. *Journal of Computational Biology* **18**, 305–321.

HANSEN, K. D., BRENNER, S. E. AND DUDOIT, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**, e131.

HAWKINS, R. D., HON, G. C. AND REN, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews. Genetics* **11**, 476–486.

HOWARD, B. E. AND HEBER, S. (2010). Towards reliable isoform quantification using RNA-Seq data. *BMC Bioinformatics* **11** (Suppl 3), S6.

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.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. AND HAUSSLER, D. (2002). The human genome browser at UCSC. *Genome Research* **12**, 996–1006.

LANGMEAD, B., TRAPNELL, C., POP, M. AND SALZBERG, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25.

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, J., JIANG, H. AND WONG, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology* **11**, R50.

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.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. AND WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628.

OSHLACK, A., ROBINSON, M. D. AND YOUNG, M. D. (2010). From RNA-Seq reads to differential expression results. *Genome Biology* **11**, 220.

PACHTER, L. (2011). Models for transcript quantification from RNA-Seq. *Arxiv*. http://arxiv.org/abs/1104.3889.

PEPKE, S., WOLD, B. AND MORTAZAVI, A. (2009). Computation for ChIP-Seq and RNA-Seq studies. *Nature Methods* **6** (11 Suppl), S22–S32.

PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J. B., STEPHENS, M., GILAD, Y. AND PRITCHARD, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772.

PRUITT, K. D., TATUSOVA, T., KLIMKE, W. AND MAGLOTT, D. R. (2009). NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Research* **37** (Database issue), D32–D36.

RICHARD, H., SCHULZ, M. H., SULTAN, M., NURNBERGER, A., SCHRINNER, S., BALZEREIT, D., DAGAND, E., RASCHE, A., LEHRACH, H., VINGRON, M., HAAS, S. A. *and others* (2010). Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research* **38**, e112.

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.

SALZMAN, J., JIANG, H. AND WONG, W. H. (2011). Statistical modeling of RNA-Seq data. *Statistical Science* **26**, 62–83.

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.

TURRO, E., SU, S. Y., GONCALVES, A., COIN, L. J., RICHARDSON, S. AND LEWIN, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-Seq reads. *Genome Biology* **12**, R13.

WANG, E. T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., KINGSMORE, S. F., SCHROTH, G. P. AND BURGE, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476.

WANG, Y., LIU, C. L., STOREY, J. D., TIBSHIRANI, R. J., HERSCHLAG, D. AND BROWN, P. O. (2002). Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5860–5865.

WANG, Z., GERSTEIN, M. AND SNYDER, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

WU, Z., WANG, X. AND ZHANG, X. (2011). Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* **27**, 502–508.

YANG, E., VAN NIMWEGEN, E., ZAVOLAN, M., RAJEWSKY, N., SCHROEDER, M., MAGNASCO, M. AND DARNELL, J. E. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Research* **13**, 1863–1872.

[*Received June 23, 2011; revised January 4, 2012; accepted for publication January 6, 2012*]