BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 502–508
doi:10.1093/bioinformatics/btq696

Gene expression

Advance Access publication December 17, 2010

Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq

Zhengpeng Wu[†], Xi Wang[†] and Xuegong Zhang[*]

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: RNA-Seq technology based on next-generation sequencing provides the unprecedented ability of studying transcriptomes at high resolution and accuracy, and the potential of measuring expression of multiple isoforms from the same gene at high precision. Solved by maximum likelihood estimation, isoform expression can be inferred in RNA-Seq using statistical models based on the assumption that sequenced reads are distributed uniformly along transcripts. Modification of the model is needed when considering situations where RNA-Seq data do not follow uniform distribution.

Results: We proposed two curves, the global bias curve (GBC) and the local bias curves (LBCs), to describe the non-uniformity of read distributions for all genes in a transcriptome and for each gene, respectively. Incorporating the bias curves into the uniform read distribution (URD) model, we introduced non-URD (N-URD) models to infer isoform expression levels. On a series of systematic simulation studies, the proposed models outperform the original model in recovering major isoforms and the expression ratio of alternative isoforms. We also applied the new model to real RNA-Seq datasets and found that its inferences on expression ratios of alternative isoforms are more reasonable. The experiments indicate that incorporating N-URD information can improve the accuracy in modeling and inferring isoform expression in RNA-Seq.

Contact: zhangxg@tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 6, 2010; revised on November 25, 2010; accepted on December 11, 2010

1 INTRODUCTION

Applying next-generation high-throughput sequencing technologies to transcriptomes (RNA-Seq) makes it possible to precisely measure the abundances of transcripts. This technology has been widely investigated on the study of gene expressions (Marioni *et al*., 2008; Nagalakshmi *et al*., 2008), splice variants (Pan *et al*., 2008; Wang *et al*., 2008; Wilhelm *et al*., 2008), novel transcripts discovery (Mortazavi *et al*., 2008), RNA sequence polymorphism (Cloonan *et al*., 2008) and chimeric transcripts (Zhang *et al*., 2010) in recent literature. The ability of RNA-Seq to count sequenced reads along transcripts gives digital measurement of transcription levels, and is powerful for identifying and measuring the amount of transcribed isoforms of alternative splicing genes (Wang *et al*., 2008).

Changes of isoforms' expression levels in many genes are of functional importance in particular biological processes. For example, switching between major and minor isoforms of genes plays a crucial role in mouse muscle myogenesis (Trapnell *et al*., 2010); isoform expression changes have been shown to be highly related to the development of many complex diseases, such as Lewy bodies (Beyer *et al*., 2008; Humbert *et al*., 2007) and progressive supranuclear palsy (Chambers *et al*., 1999); studies also obtained the evidence that the expression ratio of two alternative isoforms from human progesterone receptor (PR) gene could result in different outcomes of breast cancer treatment (Cork *et al*., 2008). Therefore, correctly estimating isoform expression levels is becoming increasingly important for understanding complicated biological mechanisms.

In RNA-Seq, under the assumption that the sequenced reads are sampled independently and uniformly on all transcripts in the sample, it is straightforward to model the distribution of read counts of exons as a Poisson distribution (Mortazavi *et al*., 2008). Using a uniform read distribution model (URD model for short), Jiang and Wong (2009) proposed a maximum likelihood estimate (MLE) method to infer isoform expression levels. However, some studies show that there are substantial biases in high-throughput sequencing data (Dohm *et al*., 2008), which may cause non-uniformity of read distributions in RNA-Seq data (Mortazavi *et al*., 2008). Recent literatures also pointed out that hexamer priming and local sequence content can affect the sequencing preference (Hansen *et al*., 2010; Li,J. *et al*., 2010). In our investigations, we observed the distributions of RNA-Seq reads in many datasets are not uniform. Some datasets have a strong bias toward the 3′ ends of transcripts, which means that the 3′-ends usually have more reads sequenced. This phenomenon may be caused by some pre-sequencing procedures such as the poly(A)-selection step in sample preparation (see Section 4 for details).

In such situations, the accuracy of isoform expression inference based on the uniformity assumption will deteriorate. Thus, the motivation of this study is to develop a method that takes the non-uniformity of read distribution into consideration to improve isoform expression inference. Recently, several other groups also realized the importance of this problem and incorporated non-URD (N-URD) information into uniform distribution models in different ways (Howard and Heber, 2010; Li,B. *et al*., 2010). In our study, we

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

introduced two types of curves to depict the tendency of sequencing biases, namely the global bias curve (GBC) and the local bias curves (LBCs). The GBC applies to all genes in a dataset and the LBCs are specific to particular genes. The curves are estimated from the read distribution across the measured transcripts. We incorporated the GBC and LBCs with the URD model and constructed N-URD models for isoform expression inference. A series of simulation experiments were conducted to study the performance of the proposed method and to compare with state-of-the-art methods. Results show that a significant gain can be achieved in the accuracy of isoform expression inference. We applied the N-URD models on several real RNA-Seq datasets, which also illustrated the advantage of the proposed method.

2 METHODS

2.1 Notations

Assume gene G has n exons of lengths $(l_1,\ldots,l_n)$. It has m isoforms with expression levels $(\theta_1,\ldots,\theta_m)$ in an experiment. From RNA-Seq data, we have a set of reads of counts $(x_1,\ldots,x_n)$ on this gene, where x_j is the number of reads mapped onto the j-th exon of gene G. The task is to infer the expression levels $(\theta_1,\ldots,\theta_m)$ from observed read counts and the known gene structure (the exon composition of each isoform). We use an indicator matrix $(a_{ij})_{m\times n}$ to represent the gene structure, where $a_{ij}=1$ or $a_{ij}=0$ indicates that the j-th exon is included or excluded in the i-th isoform, respectively. If an exon appears in two (or more) isoforms as different lengths (e.g. the case of alternative $5'$ end or $3'$ end exons), we split the exon into several parts and treat each part as an exon. Similarly, when we consider reads mapped to exon junctions (junction reads), we can use the flanking exon segments of a junction to form a ‘pseudo-exon’ and count it as an extra exon in the inference. In the experiments, we did on data of ultra-short reads, we did not consider junction reads for simplicity as they compose only a very small part in the data.

2.2 Bias curves

We propose two types of bias curves to characterize the distribution of reads along a transcript. The GBC represents the general tendency of read distribution for the whole transcriptome, and the LBCs depict gene-specific read distributions. The curves reflect the relative read distribution bias from the $5'$ end to the $3'$ end of a gene.

We use genes with only single isoforms to estimate the GBC. This is because the read distribution of the gene with multiple isoforms is affected by the gene structure and may not reflect the general tendency of read distribution along the gene. We extracted all single-isoform genes according to the NCBI Reference Sequence (RefSeq) gene annotation (downloaded from http://genome.ucsc.edu). Because of the uncertainty in lowly expressed genes, we filter out genes with too few (<100) reads. In order to avoid the impact of local fluctuation of read distribution, we divide each gene into a small number (e.g. 10) of bins and get a bar chart or histogram of the read counts across the bins. The histograms are then normalized to mean 1, and then averaged among the number of studied single-isoform genes. This gives us the GBC, which reflects general sequencing efficiency at different relative locations of a gene in the dataset. Usually, we illustrate the GBC in the form of curves instead of bar charts. The procedure for calculating the GBC is illustrated in Figure 1. The GBC depicts the global read distribution bias which may be caused by the experimental protocol. It is consistent between technical replicates in the data we studied (Section 3).

It is also observed that read distribution of a specific gene can have its own pattern that deviates from uniform distribution and also from the GBC. This kind of bias may be due to local sequence features such as GC-content variation along transcripts and simple sequence repeats in the genome, which may affect the hexamer priming efficiency (Hansen *et al.*, 2010) and read mappability (Rozowsky *et al.*, 2009). We introduced the LBCs to describe gene-specific non-uniform distributions.

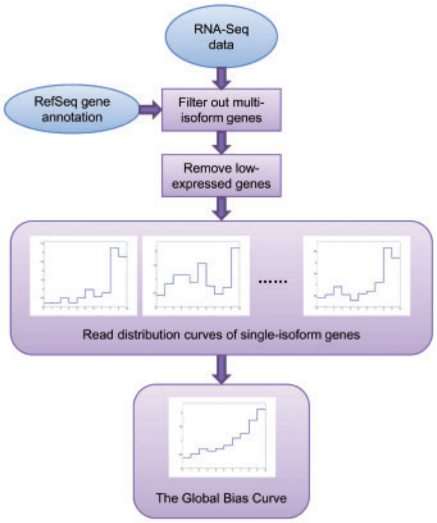
RNA-Seq data
RefSeq gene annotation
Filter out multi-isoform genes
Remove low-expressed genes
......
Read distribution curves of single-isoform genes
The Global Bias Curve

Fig. 1. The flowchart of the calculation of the GBC. The main steps include: (1) collect single-isoform genes according to known gene annotation; (2) filter out genes which have too few mapped reads (read-count <100); (3) calculate the 10-bin histograms of read positions for each gene and normalize the mean of each curve to be 1; (4) finally, the average of these bias curves is the GBC.

Read distribution of a gene with multiple isoforms is the mixture of read distributions of all its expressed isoforms. So a LBC could only be an approximate description of trend of read distribution along the gene region. We define for each gene a step function which is constant on each exon. Suppose the expression level of the i-th isoform is θ_i, we assign the value of the step function on the j-th exon as $x_j/\,(l_j\sum_{i=1}^{m}\theta_i a_{ij}),\, j=1,2,\ldots,n$, which describes the read count normalized by the exon length and isoform occurrences weighted by their expression levels. The reads on an exon also may not distribute uniformly. This can be considered by splitting an exon into multiple ‘sub-exons’ when there are sufficient reads for the exon. In our experiments, we take an exon as the basic unit for estimating the LBCs. The LBC of a gene is got by normalizing the step function to be of mean 1. At the beginning of expression inference, the isoform expression levels are unknown, so we assume expression levels of all isoforms of a gene are identical $(1,1,\ldots,1)_{1\times m}$ in the calculation of the LBC. After inferring isoform expression with this initial LBC, we can further update the LBC using the inferred isoform expression levels θ_i, and sequentially iterate this procedure. The details of this iteration procedure will be illustrated in details in Section 2.4. Figure 2 gives the conceptual flowchart for calculating the LBCs.

We use the bias curves to compensate for read-count variations caused by such biases in inferring isoform expression. For this purpose, we partition the bias curves into several segments corresponding to exons of the gene. The mean value of the curves on each segment is calculated as the weighting factor for the corresponding exon. Figure 3 illustrates this procedure. In the following sections, we will show the strategy to incorporate such weights into the statistical framework of isoform expression inference.

2.3 The URD model

In the URD model used by Jiang and Wong (2009), each observation x_j is assumed to be a random variable following a Poisson distribution with parameter λ_j. For example, λ_j for the j-th exon is $\lambda_j=l_j w\sum_{i=1}^{m}a_{ij}\theta_i$, where w is the total number of mapped reads in the RNA-Seq data and a_{ij} is the

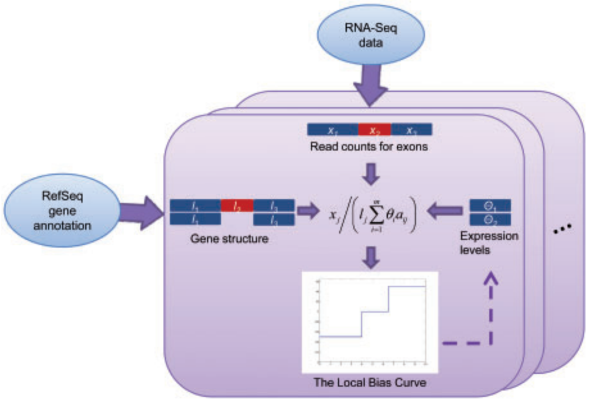

Fig. 2. The flowchart of the calculation of the LBC. Given the gene structure and RNA-Seq data, assuming the isoform expression levels are known, LBC describes the exon read counts normalized by total weighted-length of the exons appearing in all isoforms.

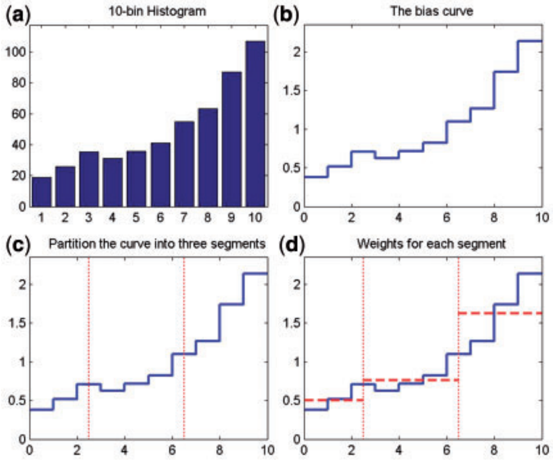

Fig. 3. Illustration of the usage of bias curve. (**a**) A gene is divided into 10 bins from its 5′ to 3′ end, and the read counts on each bin are summarized in a histogram. (**b**) The histogram is normalized to have mean value of 1 and is represented in curve form. (**c**) To utilize the bias curve in isoform expression inference, we first partition the curve into a number of parts, which equals to the number of exons in the studied isoform. The three parts (separated by red dotted lines) shown here is for a 3-exon isoform with length ratio 25:40:35. (**d**) The average value of the curve in each partition is calculated (red dashed lines). The values are used as weights for the corresponding exons in expression inference.

elements of the gene structure indictor matrix. The corresponding likelihood function is defined as

$$L(\Theta|x_j)=\frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!}.$$

Further assuming the independence of x_j's, for a gene G, the joint log-likelihood function can be written as

$$\log\left(L\left(\Theta|x_1,x_2,\ldots,x_n\right)\right)=\sum_{j=1}^{n}\log\left(\frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!}\right).$$

Inserting the $\lambda_j=l_jw\sum_{i=1}^{m}a_{ij}\theta_i$ for each exon, we have

$$\log\left(L\left(\Theta|x_1,x_2,\ldots,x_n\right)\right)=-w\sum_{j=1}^{n}\sum_{i=1}^{m}l_ja_{ij}\theta_i$$
$$+\sum_{j=1}^{n}x_j\log\left(l_jw\sum_{i=1}^{m}a_{ij}\theta_i\right)-\sum_{j=1}^{n}\log\left(x_j!\right).$$

Based on above log-likelihood function, maximum likelihood estimation can find the optimal parameters which give the inference of isoform expression levels. Further, taking derivatives for each θ_i, we get

$$\frac{\partial\log\left(L\left(\Theta|x_1,x_2,\ldots,x_n\right)\right)}{\partial\theta_i}=-w\sum_{j=1}^{n}l_ja_{ij}+\sum_{j=1}^{n}\frac{x_ja_{ij}}{\sum_{i=1}^{m}a_{ij}\theta_i}$$

Due to the convexity of above optimization problem, the gradient descending method can be used to find the solution (Jiang and Wong, 2009).

2.4 The N-URD models

Upon the above URD model, we propose N-URD models by substituting the indicator matrix (a_{ij}) with a weighted indicator matrix (b_{ij}) whose elements are non-negative real numbers calculated from the bias curves. We can rewrite the log-likelihood function as

$$\log\left(L\left(\Theta|x_1,x_2,\ldots,x_n\right)\right)=-w\sum_{j=1}^{n}\sum_{i=1}^{m}l_jb_{ij}\theta_i$$
$$+\sum_{j=1}^{n}x_j\log\left(l_jw\sum_{i=1}^{m}b_{ij}\theta_i\right)-\sum_{j=1}^{n}\log\left(x_j!\right).$$

Rather than the 0–1 indicator matrix (a_{ij}), the weighted indicator matrix (b_{ij}) not only represents the gene structure information, but also gives weights to the non-zero elements according to the bias tendency of corresponding exons. It is notable that changing the indicator matrix to the weighted indicator matrix will not change the convexity of the optimization problem. Different choice of the weighted matrix will result in different N-URD models.

Given a bias curve, say, the GBC, we assign weights to each exon in the studied isoform following the strategy in Figure 3. We repeat this procedure for every isoform in the gene, and obtain the weighted indicator matrix (G_{ij}), whose non-zero element represents the relative weight of j-th exon in the expression of i-th isoform. Here, (G_{ij}) indicates the weighted matrix is got from the GBC. We call the N-URD model with $(b_{ij})=(G_{ij})$ as the GN-URD model.

Similarly, we can get the weighted indicator matrix (L_{ij}) starting with a LBC. Letting $(b_{ij})=(L_{ij})$ results in the LN-URD model. To make use of the combination of the GBC and LBC, we can let $(b_{ij})=(G_{ij})\alpha+(L_{ij})(1-\alpha)$ and get the MN-URD model. In this study, we chose $\alpha=0.5$ to illustrate the property of the MN-URD model. Besides these three models, we further study two additional models which use iterative calculation of the LBC in the MN-URD model. We denote them by 1-M and 5-M which mean the number of iterations is 1 and 5, respectively. The iterative procedure is:

(1) Set $(\hat{\theta}_1,\ldots,\hat{\theta}_m)=(1,1,\ldots,1)$ and STEP $=0$;

(2) Calculate the LBC using $x_j/\left(l_j\sum_{i=1}^{m}\hat{\theta}_ia_{ij}\right)$, and then get $\left(b_{ij}\right)=\left(G_{ij}\right)\alpha+\left(L_{ij}\right)\left(1-\alpha\right)$;

(3) Conduct expression inference based on (b_{ij}) using MN-URD, and get the estimated expression values $(\hat{\theta}_1,\ldots,\hat{\theta}_m)$;

(4) STEP $=$ STEP $+1$, if STEP exceed the maximum number, exit; otherwise go to Step 2.

2.5 Simulation design

As there is no benchmark data with known isoform expressions, we designed a series of simulation experiments to study the performance of the proposed method. In each experiment setting (with fixed numbers of isoforms and exons), a number of simulated genes were generated in a random manner. The indicator matrices that specify gene structure were created first. We set the proportion of 1's in the elements of indicator matrices to be consistent with the RefSeq gene annotation (~80%). The lengths of exons of the simulated genes were randomly sampled from the exon lengths in the RefSeq gene annotation. In order to simulate transcriptome data with similar read distribution properties as real data, for each simulated isoform, we randomly sampled the expression level (RPKM) and the corresponding LBC from the RPKMs and their corresponding LBCs of single-isoform genes in real RNA-Seq data. For each isoform, we assigned weights of exons according to the corresponding LBC (similar to Fig. 3c and d). The reads count for each exon of the gene is determined through sampling from Poisson models.

We studied genes with numbers of isoforms varying from 2 to 5 and numbers of exons from 6 to 13. For each setting, we generated 1000 different gene structures randomly. We applied the original URD model and the proposed N-URD models to the simulated data and compared their performances.

2.6 RNA-Seq datasets

The proposed N-URD model was applied to two transcriptome RNA-Seq datasets. We also used these data to provide the information (single-isoform gene expression levels and LBCs) for generating our simulation data. The dataset by Marioni *et al.* (2008) is composed of about 120 millions reads from the human liver and kidney tissues. The dataset by Pan *et al.* (2008) consists of transcriptome data from diverse normal human tissues each of which have 17–32 millions reads. We downloaded both of them from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra). The read lengths of the two datasets are 36 and 32 bp, respectively. For brevity, we refer to these two datasets as Marioni data and Pan data, respectively. We used TopHat (Trapnell *et al.*, 2009) to map reads to the human genome assembly UCSC hg18 (NCBI build36) with up to two mismatches allowed. Reads mapped to multiple genomic loci were excluded. We observed that RNA-Seq data of different tissues from the same laboratory share large similarity on the global pattern of read distribution, so we only took the data of one tissue in each dataset as an example for the presentation and demonstration.

3 RESULTS

3.1 Bias curves describe the N-URDs

We first calculated the GBCs in the two real RNA-Seq datasets. The GBCs for each lane of the Marioni data are shown in Figure 4, and that for the Pan data is shown in Supplementary Figure S1. Each curve in Figure 4 corresponds to a technical replicate of the liver tissue sample in Marioni RNA-Seq data. All curves show strong distribution bias toward 3′ ends of transcripts. We also see obvious distribution biases on single-isoform genes (e.g. Supplementary Fig. S2). The high correlation coefficient (averaged $R^2 = 0.996$) between the GBCs of the technical replicates indicates that there are consistent distribution patterns in the same experiment. Comparing Figure 4 with Figure S1, the different shapes between GBCs from the Marioni data and from the Pan data indicates the existence of batch-effect between experiments.

We divided the whole set of single isoform genes in the Marioni data into several subgroups according to lengths of genes and compared the corresponding GBCs. The results are shown in Supplementary Figure S3. We observed very similar GBC patterns between the groups. Therefore, the read distribution bias described by the GBC reflects the consistent pattern that holds for genes with different lengths.

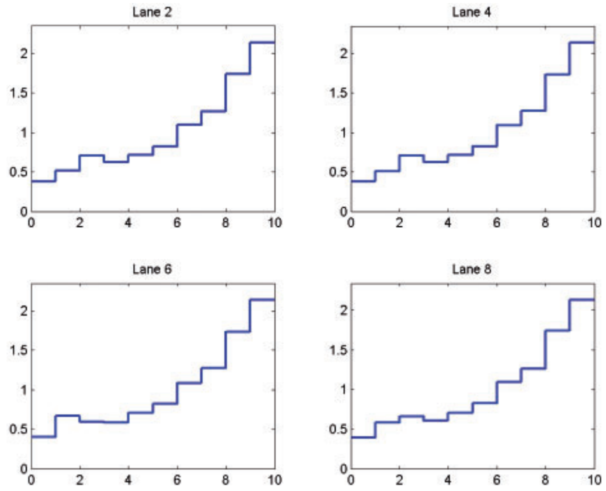

Fig. 4. The GBCs obtained in the Marioni data. The four curves represent the four technical replicates of the liver tissue sample.

3.2 Experiments on simulated data

We applied the proposed N-URD models and the original URD model to simulation datasets generated according to expression patterns in the Marioni data and Pan data. Each method gives an estimation of the expression levels of all isoforms on each simulated data. In a gene, the estimated values of isoforms can be summarized as an expression vector or a frequency vector for relative expression. There are several ways for quantitatively comparing two frequency vectors. However, as the importance of different isoforms varies and has not been well understood in most cases, such measurements are not straightforward for biological explanations. In most biological studies, it is most important to identify the major isoforms (the isoforms with highest expression among their alternatives) of alternatively spliced genes in the studied sample. Therefore, we use a *major isoform recovery rate* or MIRR to assess the estimation of the isoform expression, defined as the percentage of genes whose major isoforms are correctly identified.

Table 1 summarizes the MIRRs achieved by all compared methods in 1000 simulated genes for each (m,n) setting. These simulations were based on the Marioni data. Results with simulations based on the Pan data are provided in Supplementary Table S1. We can see that the N-URD model with different implementations all show improvement over the URD model. On average, MN-URD, 1-M and 5-M improve the MIRR by 16.7, 16.9 and 17.0%, respectively, on the Marioni simulation data, respectively. We can also observe that with the increase of the number of isoforms (m), the MIRR given by both the URD and the N-URD models goes down, because more isoforms induce more uncertainty in the inferences and the expression level of the major isoform is less different from those of minor isoforms. The advantage of N-URD models is more obvious in those more complicated situations. All methods give better performance when there are more exons (n) in genes with the same number of isoforms. Overall, different implementations of the N-URD model perform similarly but the MN-URD and 1-M, 5-M methods perform better in most cases, demonstrating the benefits of incorporating the global and local distribution information.

Table 1. Summary of the MIRR achieved by different models in simulation studies based on the Marioni data

(m,n)	URD	GN-URD	MN-URD	LN-URD	1-M	5-M
2,6	72.2	**77.9**	77.7	**77.9**	77.7	77.7
2,7	77.1	**82.0**	**82.0**	80.8	**82.0**	**82.0**
2,8	77.5	81.2	**82.0**	81.8	81.8	81.8
2,9	76.2	81.7	81.4	**81.9**	81.5	81.4
2,10	78.4	82.0	82.3	82.1	82.3	**82.4**
2,11	79.7	**82.8**	**82.8**	82.3	**82.8**	**82.8**
2,12	79.0	81.4	82.7	**83.0**	82.8	82.8
2,13	79.0	82.2	84.0	**84.1**	84.0	**84.1**
3,6	60.4	**70.5**	69.5	66.7	69.9	70.0
3,7	58.9	66.5	68.4	67.2	68.5	**68.8**
3,8	64.2	70.5	71.1	70.5	71.4	**71.7**
3,9	60.3	69.3	70.4	68.4	**71.3**	**71.3**
3,10	67.1	72.3	**73.6**	72.4	73.0	72.9
3,11	64.3	71.0	**71.8**	71.3	71.3	71.7
3,12	67.6	75.3	76.1	73.4	**76.2**	**76.2**
3,13	66.8	74.6	76.1	75.3	75.7	**76.2**
4,6	52.5	**64.5**	62.4	55.8	63.2	64.0
4,7	49.8	60.8	61.7	57.0	61.3	**61.8**
4,8	55.0	65.8	**67.2**	62.6	66.6	66.8
4,9	53.7	64.4	64.3	60.7	**65.2**	64.6
4,10	57.4	66.0	**68.3**	65.2	68.0	68.1
4,11	54.3	65.7	67.0	62.8	**67.3**	67.2
4,12	60.2	69.0	69.8	67.5	70.0	**70.1**
4,13	59.6	68.6	69.5	65.8	69.6	**70.0**
5,6	44.2	**59.4**	57.3	47.5	58.2	57.9
5,7	44.6	58.2	**59.4**	50.3	58.9	58.5
5,8	48.7	62.5	62.3	55.0	62.9	**63.3**
5,9	49.4	64.3	**64.9**	57.1	64.8	64.5
5,10	50.9	63.6	64.9	60.9	**65.3**	64.4
5,11	49.5	61.2	62.3	58.6	**62.5**	**62.5**
5,12	53.6	62.8	63.1	59.2	**63.9**	63.5
5,13	51.9	64.2	66.0	63.9	66.6	**67.0**

The first column indicates the parameters in the simulation: m is the number of isoforms, and n the number of exons. For each (m,n) pair, we generated 1000 genes to calculate the MIRRs. The columns 2–7 list the MIRRs under each model. The figures in bold indicate the largest MIRR in each row.

Besides the MIRR, we also used a more comprehensive measurement to compare the estimated expression vector with the true expression vector. We defined a *difference score* or DS for this purpose. If the true expression levels of m isoforms are $(\theta_1,...,\theta_m)$ and the estimates are $(\hat{\theta}_1,...,\hat{\theta}_m)$, the DS is defined as

$$DS = 100\sum_{i=1}^{m}\left|\frac{\theta_i}{\sum_{j=1}^{m}\theta_j}-\frac{\hat{\theta}_i}{\sum_{j=1}^{m}\hat{\theta}_j}\right|.$$

The smaller DS reflects the more accurate inference.

Supplementary Tables S2 and S3 summarize the DSs of different models on simulation data generated from the Marioni data and the Pan data with different parameter (m, n) settings. We can see that the N-URD models reduce the DS by at least 24% and 10% comparing to the URD model on the two simulation datasets, respectively. Similar to the observation with MIRRs, MN-URD and the iterative methods (1-M and 5-M) achieve the smallest averaged DS in most cases. One-sided paired t-tests show that the improvements of the N-URD model over the original URD model are significant (Supplementary Tables S4 and S5).

We also generated another set of simulation data with gene structures from RefSeq gene annotations. These simulation results are consistent with the above simulation studies (see Supplementary Table S6).

3.3 Application on real RNA-Seq data

We applied the both the URD model and the MN-URD model on the liver transcriptome RNA-Seq data from Marioni *et al.* (2008). For the 3946 genes annotated with multiple isoforms that have at least 10 reads per gene in the data, the identified major isoforms by URD and MN-URD are consistent on 3554 (90.1%) genes. We calculated the DSs between the estimates of the URD and MN-URD models for each gene with multiple isoforms. The histograms of the DSs are shown in Supplementary Figures S4 and S5. Among the genes with big DSs (DS >30 for two-isoform genes, or DS >60 for three-isoform genes), a few examples are shown in Figure 5 with the read distribution along the transcripts. The corresponding inference results by the URD and MN-URD models are listed in Table 2.

As an example, the difference of the URD and MN-URD models can be clearly seen in gene SLCO2B1. The gene has three isoforms in the annotation, but the isoform NM_007256 is inferred as not expressed by both methods. For the other two isoforms, the URD model infers that the NM_001145212 is the major isoform whereas the MN-URD shows that the NM_001145211 is the major one. We can grasp the possible reasons for this discrepancy by looking at the details of the read distribution along this gene in Figure 5. It can be seen that the density of reads falling in the leftmost four exons of NM_001145211 is not comparable with that of the right side exons. Because these four exons are included in NM_001145211 but not in NM_001145212, the read density of these four exons would reflect the expression level of NM_001145211. In the URD model, the estimation of expression levels of the two isoforms is dominated by strong signal near the 3′ end. This makes the contribution of the alternative exons very small in the estimation and causes inaccurate inference of the isoform expression. In the MN-URD model, the strong biased signal near the 3′ end was weighted down, which enables more accurate inference of the isoform expression from the exons away from the 3′ end. Similar explanations can be found in the other two examples and many other genes.

4 DISCUSSION

N-URD along a transcript is widely observed in many RNA-Seq datasets. We introduced bias curves to describe such distribution patterns in each dataset and in each gene, and proposed to modify the uniform model with these distribution informations. Experiments show that the proposed non-uniform models can give better estimation of isoform expression levels and identify major isoforms more accurately.

Biological reasons of the non-uniformity in the read distribution are not fully understood yet. The degradation of mRNA may play a main role for the global distribution pattern. As reported in literature, mRNA degradation can initiate from either the 5′ end or 3′ end of mRNAs (Beelman and Parker, 1995). In mRNA studies, usually a ploy(A) selection step is adopted. This purification also removes the RNA molecules decaying from the 3′ ends, but not those decaying

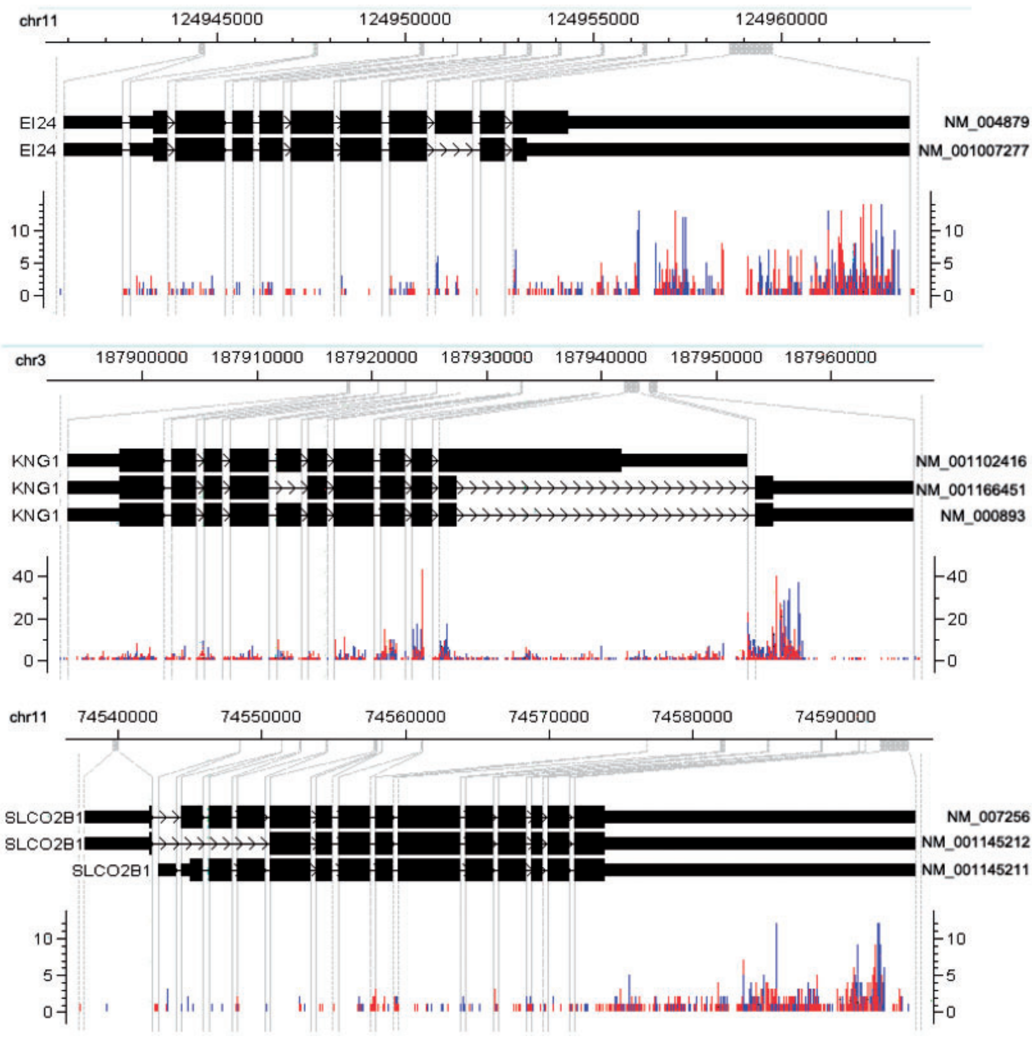

Fig. 5. Visualization of RNA-Seq reads of human genes EI24, KNG1 and SLCO2B1 in the Marioni data with CisGenome Browser (Ji *et al.*, 2008). For each panel, the three horizontal tracks in the picture from top to bottom are: genomic coordinates, gene structure where exons are magnified for better visualization and the reads falling into every two genomic coordinates, where the red and blue bars represent numbers of reads on the forward and the reverse strand, respectively.

from the 5′ ends. As a result, if the RNA-Seq library is constructed from total RNA with partially degraded RNA molecules, there would be more reads toward the 3′ ends of the transcripts, and the read distribution is therefore similar to what is shown in Figure 4 and examples in Supplementary Figure S2. One the other hand, local sequence features could explain a large part of the gene-specific distribution patterns (Li,J. *et al.*, 2010). Sequence constitution such as GC-content along genes may play the most important role in amplification, while other features may lead to biased hexamer priming (Hansen *et al.*, 2010). Studying the cause of sequencing preference may help to improve the protocol of RNA-Seq, but recently published RNA-Seq data, such as the data generated by Harr and Turner (2010), also suffered from a strong global distribution bias and local biases as well.

Recently, Li,B. *et al.* (2010) and Howard and Heber (2010) also addressed the non-uniform distribution issue from different viewpoints and proposed their models and methods for gene and/or isoform expression-level estimation. Li *et al.*'s work majorly focused on solving read mapping uncertainty, but their method could accommodate the general distribution bias in the expression inference through the usage of the empirical read start position distribution (RSPD). They did not utilize the local sequencing preference specific to genes (Li,B. *et al.* 2010). Our study shows that properly combining the global and local read distribution information is an effective way to improve the expression inference. Howard and Heber proposed a linear model to integrate the non-uniformity information, but their model was based on a normal approximation rather than the Poisson distribution. This makes their model less accurate in the inference. We conducted a series of experimental comparison with these newest methods and the results are summarized in the Supplementary Materials.

For the future study, several aspects in N-URD modeling could be addressed. First, as the read length by new sequencing technologies becomes longer, there will be more junction reads crossing two

Table 2. The inferred isoform expression values of three genes by the URD and MN-URD models in the Marioni data

Gene	Isoform	URD		MN-URD	
		RPKM	%	RPKM	%
EI24	NM_004879	206.0	56.0	273.7	75.1
	NM_001007277	161.6	44.0	90.7	24.9
KNG1	NM_001102416	158.8	24.7	138.6	21.6
	NM_001166451	273.8	42.5	96.4	15.0
	NM_000893	211.4	32.8	406.3	63.4
SLCO2B1	NM_007256	0.0	0.0	0.0	0.0
	NM_001145212	65.5	58.2	16.7	15.1
	NM_001145211	47.0	41.8	94.3	84.9

or more exons. These junction reads make up ‘pseudo-exons’ and could be helpful in the inference of alternative spliced isoforms and their expression levels. The proper combination of the GBC and LBCs is another important question. In this study, we used a fixed mixture parameter (0.5) to integrate them according to simulation experiments. More adaptively strategies, such as dynamic weighting of the two bias curves during the iteration of expression inference, would give further improvement. Another issue is about the gene annotation. The current model depends on known gene annotation for the inference. Ideally, it will be more useful if the inference can be done with incomplete annotation. In such scenario, the task of inferring isoform expression level will be integrated with the task of detecting alternative splicing isoforms from RNA-Seq data.

Accurately measuring the expression of genes and their multiple isoforms is the first step in applying RNA-Seq technology to many biological investigations. With the promising results obtained by the N-URD models, we believe that the data-adaptive strategy proposed here will benefit downstream analyses such as detecting differentially expressed genes and isoforms in human diseases.

ACKNOWLEDGEMENTS

We would like to thank Prof. Fengzhu Sun and Dr. Hui Jiang for helpful discussion. We thank the anonymous reviewers for their helpful comments and suggestions.

Funding: National Natural Science Foundation of China (grant numbers 30625012 and 60721003 in parts).

Conflict of Interest: none declared.

REFERENCES

Beelman,C.A. and Parker,R. (1995) Degradation of mRNA in eukaryotes. *Cell*, **81**, 179–183.

Beyer,K. *et al.* (2008) Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in Lewy body disease. *Neurogenetics*, **9**, 163–172.

Chambers,C.B. *et al.* (1999) Overexpression of four-repeat tau mRNA isoforms in progressive supranuclear palsy but not in Alzheimer’s disease. *Ann. Neurol.*, **46**, 325–332.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.

Cork,D.M. *et al.* (2008) Alternative splicing and the progesterone receptor in breast cancer. *Breast Cancer Res.*, **10**, 207.

Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

Hansen,K.D. *et al.* (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.

Harr,B. and Turner,L.M. (2010) Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Mol. Ecol.*, **19** (Suppl. 1), 228–239.

Howard,B.E. and Heber,S. (2010) Towards reliable isoform quantification using RNA-SEQ data. *BMC Bioinformatics*, **11** (Suppl. 3), S6.

Humbert,J. *et al.* (2007) Parkin and synphilin-1 isoform expression changes in Lewy body diseases. *Neurobiol. Dis.*, **26**, 681–687.

Ji,H. *et al.* (2008) An integrated software system for analyzing chip-chip and chip-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

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

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

Li,J. *et al.* (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R50.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

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

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

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

Rozowsky,J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

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.

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

Wilhelm,B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

Zhang,G. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.*, **20**, 646–654.