BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 8 2009, pages 1026–1032
doi:10.1093/bioinformatics/btp113

Gene expression

Statistical inferences for isoform expression in RNA-Seq

Hui Jiang[1] and Wing Hung Wong[2,*]

[1]Institute for Computational and Mathematical Engineering and [2]Department of Statistics, Stanford University, Stanford, CA 94305, USA

Received on October 12, 2008; revised and accepted on February 22, 2009

Advance Access publication February 25, 2009

Associate Editor: David Rocke

ABSTRACT

Summary: The development of RNA sequencing (RNA-Seq) makes it possible for us to measure transcription at an unprecedented precision and throughput. However, challenges remain in understanding the source and distribution of the reads, modeling the transcript abundance and developing efficient computational methods. In this article, we develop a method to deal with the isoform expression estimation problem. The count of reads falling into a locus on the genome annotated with multiple isoforms is modeled as a Poisson variable. The expression of each individual isoform is estimated by solving a convex optimization problem and statistical inferences about the parameters are obtained from the posterior distribution by importance sampling. Our results show that isoform expression inference in RNA-Seq is possible by employing appropriate statistical methods.

Contact: whwong@stanford.edu

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

1 INTRODUCTION

Recently, ultra high-throughput sequencing of RNA (RNA-Seq) has been developed as an approach for transcriptome analysis in several different species such as yeast (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008), *Arabidopsis* (Lister *et al.*, 2008), mouse (Cloonan *et al.*, 2008; Mortazavi *et al.*, 2008) and human (Marioni *et al.*, 2008; Pan *et al.*, 2008; Sultan *et al.*, 2008; Wang *et al.*, 2008). By obtaining tens of millions of short reads from the transcript population of interest and by mapping these reads to the genome, RNA-Seq produces digital (counts) rather than analog signals and offers the chance to detect novel transcripts. Because of these desirable features, not shared by qRT-PCR or microarray-based methods, RNA-Seq is widely regarded as an attractive approach to measure transcription in an unbiased and comprehensive manner. Several protocols for RNA-Seq experiments and methods for transcript quantification have been developed (see above references). In particular, in Mortazavi *et al.* (2008), the expression level of a transcript is quantified as reads per kilobase of the transcript per million mapped reads to the transcriptome (RPKM). By normalizing the counts of reads mapped to (all the exons belonging to) a gene against the transcript length and the sequencing depth, the RPKM index makes it easy to compare expression measurements across different genes and different experiments. In mouse liver tissue samples, RNA-Seq based and exon array-based expression indexes across RefSeq genes are very well correlated (rank correlation >0.85) (Kapur *et al.*, 2008). Given that the datasets were generated by independent laboratories [RNA-Seq from Mortazavi *et al.* (2008), exon arrays from Affymetrix], this high degree of concordance gives us confidence on the quantitative accuracy of both methods for gene-level expression analysis.

In principle, the RPKM concept is equally applicable to quantify isoform expression—it is simply the counts of reads mapped to a specific isoform normalized against the isoform length and the sequencing depth. In practice, however, it is difficult to compute isoform-specific RPKM because most reads that are mapped to the gene are shared by more than one isoform. In this article, we developed a statistical model to describe how the counts mapped to the exons of the gene are related to the isoform-specific expression indexes. Based on this model, we studied the statistical methods to estimate isoform-specific expression indexes and to quantify the uncertainties in the estimates. Our method can be viewed as an extension of the RPKM concept and reduces to the RPKM index when there is only one isoform. It was found that in many cases standard inference based on the asymptotic distribution of the maximum likelihood estimate (MLE) is inadequate because the Fisher information matrix is almost singular or the MLE is near the boundary of the parameter space. To address these difficulties, we have developed a Bayesian inference method based on importance sampling from the posterior distribution.

2 THE MODEL

2.1 Notations

First we introduce the notations. Let G be the set of genes. For any gene $g \in G$, let $F_g = \{f_{g,i} | i \in [1, n_g]\}$ be the set of its isoforms, where n_g is a positive integer. Also, let $F = \{f_{g,i} | g \in G, i \in [1, n_g]\}$ be the set of all the possible isoforms of all the genes, which stands for all the different possible transcripts in the sample being sequenced. For any isoform $f \in F$, let l_f be its length, and let k_f be the number of copies of transcripts in the form of isoform f in the sample.

Based on the above notations, we know that the total length of the transcripts in the sample is $\sum_{f \in F} k_f l_f$. We model sequencing process as a simple random sampling, in which every read is sampled independently and uniformly from every possible nucleotides in the sample. Therefore, the probability that a read comes from some isoform f is $p_f = k_f l_f / \sum_{f \in F} k_f l_f$. By defining $\theta_f = k_f / \sum_{f \in F} k_f l_f$ as the expression index of isoform f in the sample, we can rewrite

*To whom correspondence should be addressed.

p_f as $p_f = \theta_f l_f$ and then we have $\sum_{f \in F} \theta_f l_f = 1$, which makes the model identifiable.

Let w be the total number of mapped reads. Given an isoform f, and a region of length l in f, because in our model the reads are sampled uniformly and independently, we know that the number of reads coming from that region, denoted by some random variable X, follows a binomial distribution with parameters w and $p = \theta_f l$. Since usually w is very large and p is very small, the binomial distribution here can be approximated well by a Poisson distribution with parameter $\lambda = w\theta_f l$. The uniform model and the Poisson approximation have been successfully used and tested in several previous RNA-Seq studies such as Mortazavi *et al.* (2008) and Marioni *et al.* (2008).

In general, for a given read we can map it to the reference sequence so that we know where it comes from. However, in cases that a gene has more than one isoform, these isoforms often share some common regions (e.g. common exons), thus making it very difficult for us to determine the true isoform a read actually comes from if that read is mapped to a common region.

As a summary, suppose that the gene structures and the sequencing reads are all given, i.e. G, F, l, w are all known, the problem left to us now is to estimate θ.

2.2 Poisson model

We deal with the problem by solving a Poisson model. For a gene g, suppose it has m exons with lengths $\mathbf{L} = [l_1, l_2, \ldots, l_m]$ and n isoforms with expressions $\mathbf{\Theta} = [\theta_1, \theta_2, \ldots, \theta_n]$, where $\mathbf{L}$ and $\mathbf{\Theta}$ are vectors. Suppose these isoforms only share exons as a whole, i.e. they either share an exon or do not share it. In cases that two isoforms share part of an exon, we can split the exon into several parts and then treat each part as an exon separately.

Suppose we have a set of observations $\mathbf{X} = \{X_s | s \in \mathcal{S}\}$, where S is an index set, and each observation $X \in \mathbf{X}$ is a random variable representing the number of reads falling into some region of interest in g. For example, reads falling into some exon, or reads falling into some exon–exon junction.

From the above observation, we know for every $X \in \mathbf{X}$ that it follows a Poisson distribution with some parameter λ. For instance, the λ for the number of reads falling into exon j is $l_j w \sum_{i=1}^{n} c_{ij}\theta_i$, where c_{ij} is 1 if isoform i contains exon j and 0 otherwise. For exon–exon junctions, the λ is $lw\sum_{i=1}^{n} c_{ij}c_{ik}\theta_i$, where l is the length of the junction region, and j and k are indices of the two exons involved in the junction being investigated.

In general, λ is a linear function of $\theta_1, \theta_2, \ldots, \theta_n$, i.e. $\lambda = \sum_{i=1}^{n} a_i\theta_i$. From the probability mass function of the Poisson distribution, we have the likelihood function

$$\mathcal{L}(\mathbf{\Theta}|x) = P(X = x|\mathbf{\Theta}) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{1}$$

For the whole set of observations $\mathbf{X} = \{X_s | s \in \mathcal{S}\}$, if the corresponding regions do not overlap then X_s's are independent and we can write the joint log-likelihood function as

$$\log(\mathcal{L}(\mathbf{\Theta}|x_s, s \in \mathcal{S})) = \sum_{s \in \mathcal{S}} \log(\mathcal{L}(\mathbf{\Theta}|x_s)) \tag{2}$$

The MLE is obtained by

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}} \log(\mathcal{L}(\mathbf{\Theta}|x_s, s \in \mathcal{S}))$$

2.3 Single-isoform case

Taking logarithm on (1), we have

$$\begin{aligned}\log(\mathcal{L}(\mathbf{\Theta}|x)) &= -\lambda + x\log\lambda - \log(x!) \\ &= -\textstyle\sum_{i=1}^{n} a_i\theta_i + x\log\left(\sum_{i=1}^{n} a_i\theta_i\right) - \log(x!)\end{aligned} \tag{3}$$

Taking derivatives, we get

$$\frac{\partial\log(\mathcal{L}(\mathbf{\Theta}|x))}{\partial\theta_j} = -a_j + \frac{xa_j}{\sum_{i=1}^{n} a_i\theta_i} \tag{4}$$

When $n = 1$, so $\mathbf{\Theta} = \theta_1$ is a real number; it is well known that $\hat{\mathbf{\Theta}} = x/a$. When x is the number of reads falling into some region of length l, we have $a = wl$ and therefore $\hat{\mathbf{\Theta}} = x/wl$, which is equivalent to the RPKM defined in Mortazavi *et al.* (2008).

The Poisson model and the maximum likelihood estimation for genes with single isoform have been used in Marioni *et al.* (2008) for detecting differences among technical replicates.

2.4 Multiple-isoform case

When $n > 1$, i.e. the gene has more than one isoform, a simple closed-form solution is no longer available. We employ numerical methods for solving the maximum likelihood estimation problem. Here we show that the model has a nice property.

PROPOSITION 1. *The joint log-likelihood function (2) is concave.*

PROOF. Since the sum of concave functions is still concave, proving the concavity of the log-likelihood function with single observation (3) suffices.

Taking second-order derivatives of (3), we get

$$\frac{\partial^2\log(\mathcal{L}(\mathbf{\Theta}|x))}{\partial\theta_j\partial\theta_k} = -\frac{xa_ja_k}{\left(\sum_{i=1}^{n} a_i\theta_i\right)^2}$$

Consider the Hessian matrix $\mathbf{H}(\mathbf{\Theta})$ where the (j,k)-th element is

$$\mathbf{H}_{jk}(\mathbf{\Theta}) = \frac{\partial^2\log(\mathcal{L}(\mathbf{\Theta}|x))}{\partial\theta_j\partial\theta_k}$$

We can write $\mathbf{H}$ as $\mathbf{H} = -d\mathbf{a}'\mathbf{a}$ where $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ is a vector and $d = x/\left(\sum_{i=1}^{n} a_i\theta_i\right)^2$ is a scalar.

When $x \geq 0$ (which is true here because x is the observed number of reads falling into some region of interest) we have $d \geq 0$, therefore $\mathbf{H}$ is negative semi-definite because for any vector $\mathbf{y} = [y_1, y_2, \ldots, y_n]$, we have $\mathbf{yHy}' = \mathbf{y}(-d\mathbf{a}'\mathbf{a})\mathbf{y}' = -d(\mathbf{ya}')(\mathbf{ya}')' = -d(\mathbf{ya}')^2 \leq 0$. The concavity of the log-likelihood function is therefore guaranteed. ■

Given the concavity of the joint log-likelihood function, we can use any optimization method to compute the maximum likelihood estimator $\hat{\mathbf{\Theta}}$, and any local maximum is guaranteed to be a global maximum.

3 STATISTICAL INFERENCES

We are interested in statistical inference methods that go beyond point estimation. For example, we would like to quantify the degree of uncertainty in our point estimates and to identify features of the parameters that are poorly (or well) estimated.

3.1 Fisher information matrix

We know that when the true parameter is not on the boundary of the parameter space, the distribution of $\hat{\boldsymbol{\Theta}}$ can be approximated asymptotically by a normal distribution with mean $\boldsymbol{\Theta}$ and covariance matrix equal to the inverse Fisher information matrix $\mathcal{I}(\boldsymbol{\Theta})^{-1}$ [see chapter 5 of van der Vaart (1998)]. For a single observation X, we know that the (j,k)-th element of the Fisher information matrix is

$$\begin{aligned}\mathcal{I}_{jk}(\boldsymbol{\Theta}) &= \mathrm{Cov}_X\left[\frac{\partial \log(\mathcal{L}(\boldsymbol{\Theta}|X))}{\partial\theta_j}, \frac{\partial \log(\mathcal{L}(\boldsymbol{\Theta}|X))}{\partial\theta_k}\middle|\boldsymbol{\Theta}\right]\\ &= -\mathrm{E}_X\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\Theta}|X))}{\partial\theta_j\partial\theta_k}\middle|\boldsymbol{\Theta}\right]\\ &= -\mathrm{E}_X\left[-\frac{Xa_ja_k}{\left(\sum_{i=1}^n a_i\theta_i\right)^2}\middle|\boldsymbol{\Theta}\right]\\ &= \frac{a_ja_k}{\sum_{i=1}^n a_i\theta_i}\end{aligned}$$

The last equation holds because the mean of a Poisson random variable with parameter $\lambda = \sum_{i=1}^n a_i\theta_i$ is λ itself.

Since in practice we do not know the true $\boldsymbol{\Theta}$, we can use the estimated $\hat{\boldsymbol{\Theta}}$ to approximate the true $\boldsymbol{\Theta}$, or alternatively, we can use the observed Fisher information matrix at $\hat{\boldsymbol{\Theta}}$

$$\mathcal{J}_{jk}(\hat{\boldsymbol{\Theta}}) = -\left.\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\Theta}|x))}{\partial\theta_j\partial\theta_k}\right|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} = \frac{xa_ja_k}{\left(\sum_{i=1}^n a_i\hat{\theta}_i\right)^2}$$

For a set of independent observations $\mathbf{X} = \{X_s | s \in \mathcal{S}\}$, we have the Fisher information matrix and the observed Fisher information matrix for the joint distribution

$$\mathcal{I}_{jk}(\boldsymbol{\Theta}) = \sum_{s\in\mathcal{S}} \mathcal{I}_{jk}^{(s)}(\boldsymbol{\Theta}) = \sum_{s\in\mathcal{S}} \frac{a_j^{(s)}a_k^{(s)}}{\sum_{i=1}^n a_i^{(s)}\theta_i}$$

$$\mathcal{J}_{jk}(\hat{\boldsymbol{\Theta}}) = \sum_{s\in\mathcal{S}} \mathcal{J}_{jk}^{(s)}(\hat{\boldsymbol{\Theta}}) = \sum_{s\in\mathcal{S}} \frac{xa_j^{(s)}a_k^{(s)}}{\left(\sum_{i=1}^n a_i^{(s)}\hat{\theta}_i\right)^2}$$

3.2 Importance sampling from the posterior distribution

When some of the θ_i's are close to zero, i.e. some isoforms are lowly expressed, the likelihood function is truncated at $\theta_i = 0$ since all the isoform expressions should be non-negative. Therefore, we have constraints $\theta_i \geq 0$ for all i. As a result, the covariance matrix estimated by the inverse of the Fisher information matrix is no longer reliable.

To handle this difficulty, our inference on the θ_i's is based on their joint posterior distribution instead of the asymptotic distribution of their MLE. This is done by the importance sampling method. [see chapter 2 of Liu (2002)]. First, we generate random samples $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(k)}$ from a proposal density and associate with each of them an importance weight $w^{(i)} = \mathcal{L}(\boldsymbol{\Theta}^{(i)})/q(\boldsymbol{\Theta}^{(i)})$, where $\mathcal{L}(\boldsymbol{\Theta}^{(i)})$ is the likelihood function at $\boldsymbol{\Theta}^{(i)}$ and $q(\boldsymbol{\Theta}^{(i)})$ is the density of the proposal distribution at $\boldsymbol{\Theta}^{(i)}$.

Using these weighted samples, we can estimate the posterior probability of any event A as

$$P(\boldsymbol{\Theta} \in A|X) \approx \frac{\sum_{\boldsymbol{\Theta}^{(i)}\in A} w^{(i)}}{\sum w^{(i)}}$$

and the posterior expectation of any function u as

$$\mathrm{E}(u(\boldsymbol{\Theta})|X) \approx \frac{\sum w^{(i)}u(\boldsymbol{\Theta}^{(i)})}{\sum w^{(i)}}$$

This allows us to make any computation necessary for statistical inferences such as estimating θ_i by computing its posterior expectation, and quantifying the uncertainty of this estimate by computing an interval around it that contains 95% of the posterior probability (95% probability interval).

In this article, the proposal density is taken to be a multivariate normal with mean vector equal to the MLE of $\boldsymbol{\Theta}$ and covariance matrix equal to a matrix modified from the inverse of the Fisher information matrix. The modification, described in more detail below, is designed to improve the conditioning of the matrix and to reduce the variance of the importance weights.

4 RESULTS

We test our model with the RNA sequencing dataset published in Mortazavi *et al.* (2008). In this dataset, three mouse tissue samples: liver, skeletal muscle and brain are sequenced on the Solexa platform. For each tissue, 60–80 million reads from two replicates are put together.

We take gene annotations from the RefSeq mouse mRNA database (mm9, NCBI Build 37) downloaded from the UCSC Genome Browser (Karolchik *et al.*, 2008). Reference sequences for all the exons and exon–exon junctions are extracted from the mouse genome (mm9, NCBI Build 37) and all the sequencing reads are mapped to the reference sequences using SeqMap, a short sequence mapping Tool (Jiang and Wong, 2008).

Among all the 19 069 RefSeq genes in the database, 1510 genes have more than one isoform. For these genes, the average number of isoforms is 2.39 and the maximum number of isoforms is 12.

4.1 Isoform expression estimation

We use coordinate-wise hill climbing for solving the optimization problem. Individual parameters are optimized in turn until convergence. We keep all the parameters greater or equal to zero during the optimization process. This method is simple, robust and also quite efficient in our experiments. For the 1510 genes which have more than one isoform, the mean of the number of iterations before convergence is 32.87 and the median is 15.

In our computations, instead of using the actual exon length l, we use the effective exon length defined as $l-r$ (where r is the read length, which is 25 in our experiments), because it is the number of possible places that a read can be mapped to in that exon. Moreover, exons that are always either present or not present at the same time (e.g. the three left-most exons in Fig. 2b) are regarded as a single (super) exon in our computations and the reads mapped to them are put together.

We define the expression of a gene as the sum of expressions of isoforms that belong to that gene. A histogram of gene expressions in liver samples is shown in Figure 1. Comparing liver and muscle

Fig. 1. Histogram of gene expressions in liver samples in the unit of RPKM. Genes are grouped into eight log-scaled bins according to their expressions. Genes are considered to be lowly (or highly) expressed if their RPKMs are below 1 (or above 100). Genes that have RPKMs between 1 and 100 are considered to be moderately expressed.

samples, 487 genes show strong differential expression (>10-fold) and are highly expressed (expression >100) in at least one of the two tissues.

A gene (Pdlim5) whose isoforms are differentially expressed is shown in Figure 2a. In brain samples, the estimated expressions for the three isoforms (from top to bottom) are 5.05, 0.42 and 0, respectively. In muscle samples they are expressed at 1.91, 238.67 and 14.89, respectively. As we can see, the first isoform is actually downregulated in muscle, although in terms of gene-level expression it shows upregulation.

4.2 Statistical inferences

For many genes, we encounter problems when computing the Fisher information matrix $\mathcal{I}$ and the observed Fisher information matrix $\mathcal{J}$, because:

(1) The term $\sum_{i=1}^{n} a_i\theta_i$ in the denominator may be zero.

(2) The matrices may be degenerated and therefore not invertible.

To solve these problems, we add a matrix $\epsilon\mathbf{I}$ (where $\mathbf{I}$ is the identity matrix) to the computed Fisher information matrix and the computed observed Fisher information matrix.

To handle the difficulties associated with ill-conditioned Fisher information matrix and the boundary effect of the parameter space, we employ importance sampling to simulate from the posterior distribution. Based on the expression vector $\hat{\boldsymbol{\Theta}}$ estimated by optimizing the likelihood function and the covariance matrix $\mathbf{C}$ estimated by the Fisher information matrix, we generate samples from multivariate normal distribution with mean $\hat{\boldsymbol{\Theta}}$ and covariance matrix $4(\mathbf{C}+\mathrm{Tr}(\mathbf{C})\mathbf{I}/10)$, where $\mathbf{I}$ is the identity matrix and $\mathrm{Tr}(\mathbf{C})$ is the trace of $\mathbf{C}$. This choice of proposal distribution takes advantage of the correlation structure estimated in $\mathbf{C}$ and therefore will be efficient in terms of sampling. For each gene, we generate 50 000 samples from the proposal distribution and re-estimate the isoform expressions, gene expression and covariance matrix based on these samples. We also estimate 95% probability intervals for isoform expressions and gene expression. Using the importance weights we can compute these intervals as intervals that contain 95% of the probability in the marginal posterior distribution of the parameters of interest.

We define a 95% probability interval to be a narrow one if its length is <10% of the value of its corresponding gene expression. In liver, 2115 genes (11% of all the 19 069 genes) have narrow intervals and 2071 of these 2155 genes have expression >10. In the 728 highly expressed genes (expression >100), 723 (99%) of them have narrow intervals. For isoform expressions, there are 121 genes (8% of all the 1510 genes with multiple isoforms) that have narrow intervals for all isoforms, and in the 49 genes with expression >100 and multiple isoforms, 42 (86%) of them have narrow intervals for all isoforms. As we can see, less uncertainties are in gene expressions than in isoform expressions. Similar results are found in muscle and brain samples.

An example of mouse gene Dbi is shown in Figure 2b. Where the two isoforms are estimated to have expressions 3.87 and 580.68, respectively, and the 95% probability intervals estimated are (2.22, 6.76) and (559.52, 601.86), respectively. As we can see in this case, as the expression gets higher, the interval gets relatively narrower which means the estimate gets more accurate.

We notice that in many cases isoform-level inference is much less precise than gene-level inference, i.e. we are more certain about gene expression than isoform expressions. For example, in Figure 2c, the two isoforms are estimated to have expressions 0.48 and 6.60 with 95% probability intervals (0.05, 3.01) and (4.20, 7.28), respectively. However, the gene expression is estimated as 7.09 with a much narrower interval (6.52, 7.84). This is because the two isoforms are distinguished only by a very short exon which makes both of their expressions very sensitive to the number of reads falling into this exon. However, the gene expression is basically determined by all the reads falling into all the exons, which is a much larger number so that it has much smaller variance. Actually, in this case the expressions of the two isoforms are highly negatively correlated.

Another example is shown in Figure 2d. It is easy for us to tell that the first isoform (from top to bottom) is not expressed because there is no read falling into the first exon (from left to right). However, it is not easy for us to tell the expressions of the other two isoforms just by examining the mapped reads. The estimated expressions for the second and the third isoforms are 1.57 and 13.58, respectively, which is consistent with the fact that the read densities in the second and the third exons are approximately at the same level. The 95% probability intervals (0.09, 3.80) and (11.51, 15.30) show the degree of uncertainty that we have in these estimates. Figure 3 shows all one- and two-dimensional marginal posterior distributions of the three parameters.

4.3 Validation with microarray data

We compare the results derived by our approach to the results derived by a microarray approach in Pan *et al.* (2004), where customized microarrays were used to investigate 3126 'cassette-type' alternative splicing (AS) events in 10 mouse tissues, with seven probes targeting each AS event. For each event in each tissue, a percent alternatively spliced exon exclusion value (%ASex) was computed. Furthermore, some selected AS events (eight were shown in the article) in 10 tissues were validated with over 200 RT-PCR experiments and with

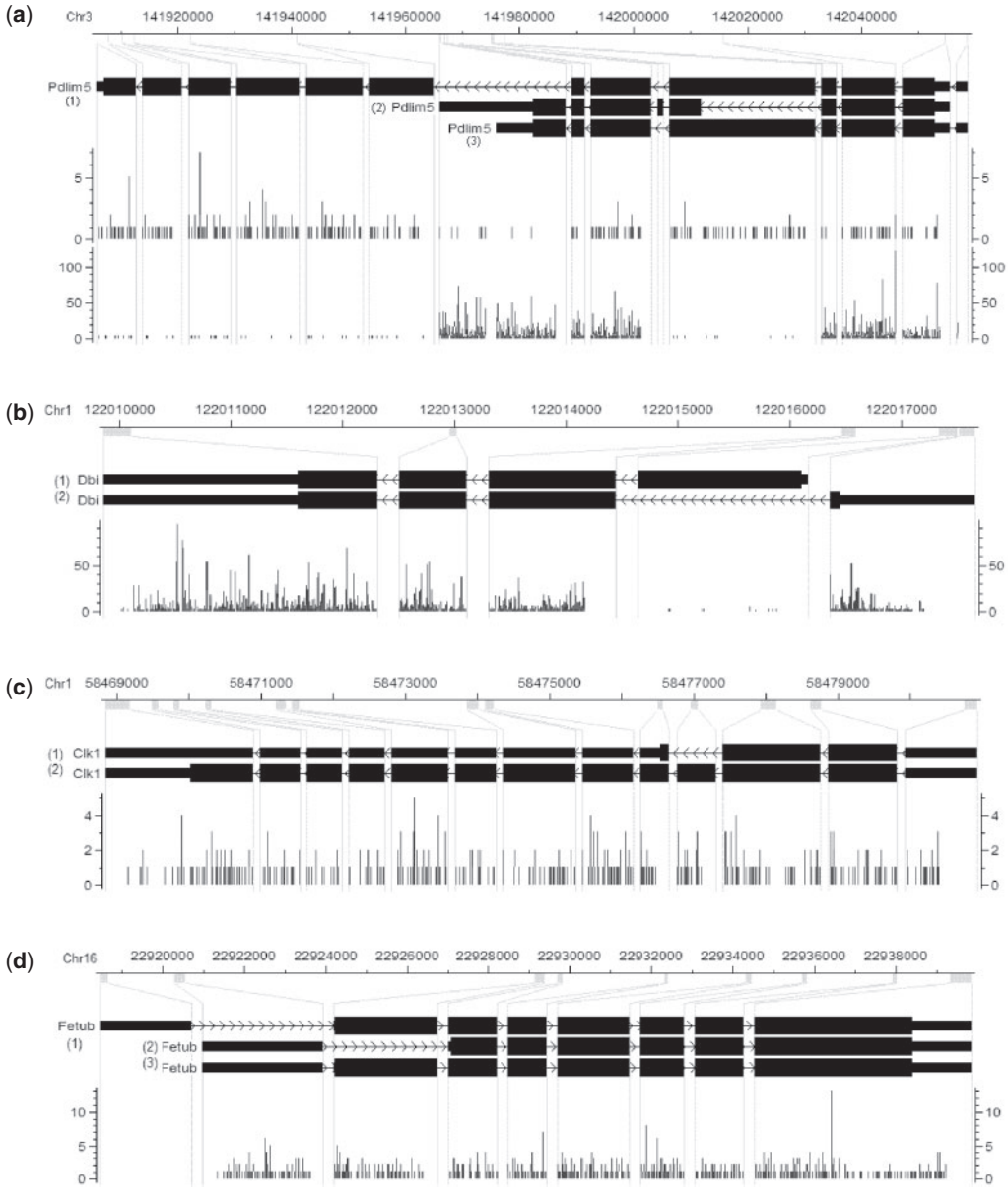

Fig. 2. (**a**) Visualization of RNA-Seq reads falling into mouse gene Pdlim5 in CisGenome Browser (Ji *et al.*, 2008). The four horizontal tracks in the picture are (from top to bottom): genomic coordinates, gene structure where exons are magnified for better visualization, the reads falling into each genomic coordinate in brain and muscle samples, where the red or blue bar represents the number of reads on the forward or reverse strand that starts at that position. Visualization of mouse genes Dbi (**b**), Clk1 (**c**) and Fetub (**d**) in brain tissue.

good consistency: Pearson's correlation coefficient (PCC) > 0.6 for all selected AS events; 60% of the selected AS events have %ASex values within ±0.15 from the values determined by RT-PCR.

For all 3126 exons [annotated with GenBank id(s) and start and end coordinates] investigated in Pan *et al.* (2004), 1193 of them can be matched to an exon in the RefSeq annotation that we are using. We then build gene models using these AS events with two isoforms (f_1 and f_2) for each gene, one isoform (f_1) with the alternatively spliced exon and the other one (f_2) without it. Afterwards, isoform and gene expression values are computed using our approach in three mouse tissues (liver, muscle and brain) using the RNA-Seq data in Mortazavi *et al.* (2008). The AS exon exclusion values are then computed as the ratios between the expression values of f_2 and the gene expression values.

We compare our computed AS exon exclusion values to the %ASex values given in Pan *et al.* (2004) on a selected set of genes. The selection criteria are:

(1) The gene should be moderately expressed, for which we use the gene expression value >5 as the cutoff, where 5 is about the median of all the gene expression values.

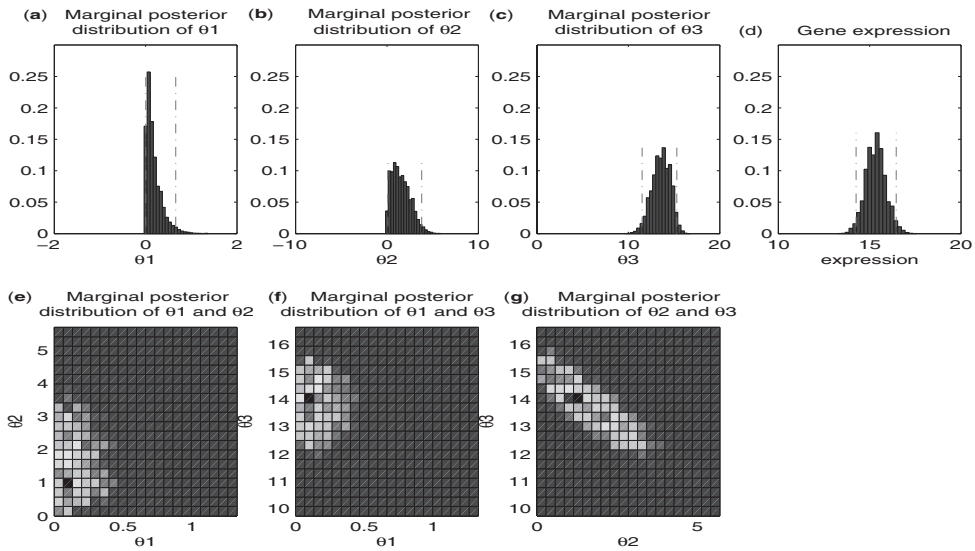

Fig. 3. Statistical inference using importance sampling for mouse gene Fetub in brain tissue (Fig. 2d). Histograms of marginal posterior distribution of θ_1, θ_2 and θ_3 are given in (**a**), (**b**) and (**c**), respectively. The gene expression is shown in (**d**) as a reference. The two red dotted vertical lines in each pictures are the boundaries for the 95% probability intervals. (**e**), (**f**) and (**g**) are the heatmaps showing marginal posterior distributions of all two-parameter combinations. We can see from the heatmaps that θ_1 is almost uncorrelated with the other two parameters, while θ_2 and θ_3 are negatively correlated.

Table 1. The number of selected AS events, the PCC between our results and the results in Pan *et al.* (2004), and the percentage of events that have differences within ±0.15 between our results and the results in Pan *et al.* (2004) for each of the three mouse tissues

Tissue	No. selected AS events	PCC	Events with differences within ±0.15(%)
Liver	472	0.48	58
Muscle	451	0.40	47
Brain	699	0.36	49

(2) The AS exon exclusion value should have a relatively small 95% probability interval, for which we use 0.5 as the cutoff, i.e. $(f_2^u - f_2^l)/g < 0.5$, where f_2^u and f_2^l are the upper and lower bounds of the 95% probability interval for the expression value of f_2, and g is the gene expression value.

Table 1 gives the number of events we selected, the PCC between our results and the results in Pan *et al.* (2004), and the percentage of events that have differences within ±0.15 between our results and the results in Pan *et al.* (2004) for each of the three mouse tissues.

We found that for some genes, the %ASex values computed in Pan *et al.* (2004) are quite large, while the values computed by our approach are close to zero. As a typical example, for gene DNAJC7 in mouse brain tissue, there are 176 reads mapped to the AS exon of length 89 nt, which is 1.98 reads per nucleotide. As a comparison, there are 3300 reads mapped to the remaining part of the gene of length 2223 nt, which is 1.48 reads per nucleotide. In addition, there are 123 reads mapped to the two exon–exon junctions that exclusively belong to f_1, and no read mapped to the exon–exon junction that exclusively belongs to f_2. All of the above evidences indicate that f_2 (the AS exon-excluding isoform) is either lowly expressed or even not present in the sample, which is concordant with the AS exon exclusion value (0.007) computed by our approach. However, the %ASex value given in Pan *et al.* (2004) is 0.433. We suspect that this type of discrepancy is caused by either errors in the annotations, or noises in the microarray probe signals. To investigate the consistency when these cases are excluded, we further filter the AS events with one additional condition: $f_2/g < 0.1$, where f_2/g is the exclusion value computed by our approach. The comparison results are given in Table 2. We found that better consistencies (in terms of PCC) are obtained in all three tissues.

Table 2. Comparison results, with refined selection criteria

Tissue	No. selected AS events	PCC	Events with differences within ±0.15(%)
Liver	228	0.60	56
Muscle	194	0.48	48
Brain	298	0.44	57

5 DISCUSSION

Our results show that isoform expression inference in RNA-Seq is possible by employing the Poisson model and appropriate statistical methods. Quantitative inferences can be drawn for a gene with one or more isoforms. When we examine the mapped reads against the annotated isoforms (e.g., Fig. 2d), we can see that our inferences are always consistent with the ones suggested by the detailed inspection. However, such gene-by-gene manual inspection cannot be scaled to the genome level. Thus, the availability of a method that can automatically extract this information should be a useful tool.

We found that exon–exon junction reads can help to reduce the 95% probability intervals because these reads fall into some isoform-specific regions and therefore provide useful information for separating the expressions of different isoforms. For instance, in mouse liver tissue with RefSeq annotations, with junction reads, the maximum relative interval [defined as $\max_{f \in F_g}(f^u - f^l)/g$] for all isoforms of a gene has an average length of 0.39 among all the genes with multiple isoforms, while the value is 0.44 without junction reads. Naturally, we believe paired-end reads and longer reads will provide even more information. However, the handling of paired-end reads will require further development.

Although uniform sampling and Poisson approximation of the sequencing reads are proved to be useful in Marioni *et al.* (2008) and also in our experiments, we do find that in some cases the reads are not truly random and uniform. This is probably caused by the technical bias in the fragmenting, priming and amplifying procedures in the sequencing experiments. In some genes, we discover exceptionally high peaks and also positive correlation in read distributions between different tissues. Although in long regions the non-random and non-uniform nature of the data may be averaged out, their effects are non-negligible when we study short regions such as single exons. We also found bias in read distributions towards the $3'$ tail and also many $3'$-UTR variants, which are important topics for future studies.

In our model we have assumed that all the isoforms for a gene are known beforehand. Currently, because of the complexity of the transcriptome and the limitations of previous experimental approaches, the isoform-level annotation is very incomplete. However, the results in this article are still valuable for two reasons. First, we can expect that many RNA-Seq datasets will be available in the near future for various cell types, making it feasible to discover most of the common isoforms. Second, we believe that isoform expression quantification and novel transcription event discovery are closely related problems and that progress towards one problem will contribute to the progress towards the other. We hope the quantitative isoform information will assist us in searching for new transcription events. For instance, one can attempt to develop a goodness-of-fit test to detect cases for which our model does not fit well, in this way we may discover gene loci where the current annotation is incomplete and where there may exist new isoforms or new AS events. Fitting the model with unknown isoforms rather than known ones may be another possible approach which requires further study. If we allow the c_{ij} (which indicates whether isoform i contains exon j) to be variable rather than fixed, our model would be able to allow exon-skipping events in the current annotations. This only requires some extra works in the estimation (e.g. using the EM algorithm to infer the unknown c_{ij}'s). However, novel isoforms consist of not only exon-skipping events but also many other events such as new exons and exon variants. Moreover, with more variabilities in the isoform structures, whether the current datasets are sufficient to solve a complex model is yet to be studied. Finally, we believe that the accuracy of the results will improve as the sequencing technology evolves and generates longer sequences with less noise and higher throughput.

ACKNOWLEDGEMENTS

We thank Barbara Wold for making the Solexa sequencing data available and Yi Xing for the Pan *et al.* (2004) reference. We also thank the three anonymous referees for their insightful comments and suggestions, which have led to an improved article.

Funding: National Institute of Health (R01HG003903, U54GM62119 and R01HG004634 to W.H.W.); California Institute for Regenerative Medicine grant (to W.H.W, partial).

Conflict of Interest: none declared.

REFERENCES

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

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. (2008) Seqmap : mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24**, 2395–2396.

Kapur,K. *et al.* (2008) Cross-hybridization modeling on affymetrix exon arrays. *Bioinformatics*, **24**, 2887–2893.

Karolchik,D. *et al.* (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.

Lister,R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Liu,J.S. (2002) *Monte Carlo Strategies in Scientific Computing*. Springer.

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.* (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell*, **16**, 929–941.

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

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

van der Vaart,A.W. (1998) *Asymptotic Statistics*. Cambridge University Press.

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.