



A structured sparse regression method for estimating isoform expression level from multi-sample RNA-seq data

L. Zhang and X.J. Liu

College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China

Corresponding author: L. Zhang

E-mail: leo.zhang@nuaa.edu.cn

Genet. Mol. Res. 15 (2): gmr.15027670

Received September 17, 2015

Accepted December 29, 2015

Published June 3, 2016

DOI <http://dx.doi.org/10.4238/gmr.15027670>

ABSTRACT. With the rapid development of next-generation high-throughput sequencing technology, RNA-seq has become a standard and important technique for transcriptome analysis. For multi-sample RNA-seq data, the existing expression estimation methods usually deal with each single-RNA-seq sample, and ignore that the read distributions are consistent across multiple samples. In the current study, we propose a structured sparse regression method, SSRSeq, to estimate isoform expression using multi-sample RNA-seq data. SSRSeq uses a non-parameter model to capture the general tendency of non-uniformity read distribution for all genes across multiple samples. Additionally, our method adds a structured sparse regularization, which not only incorporates the sparse specificity between a gene and its corresponding isoform expression levels, but also reduces the effects of noisy reads, especially for lowly expressed genes and isoforms. Four real datasets were used to evaluate our method on isoform expression estimation. Compared with other popular methods, SSRSeq reduced the variance between multiple samples, and produced more accurate

isoform expression estimations, and thus more meaningful biological interpretations.

Key words: RNA-seq; Multi-sample; Structured sparse regression; Bias curve; Isoform expression level

INTRODUCTION

High-throughput sequencing of RNA, known as RNA-seq, is a revolutionary and powerful technology for transcriptome analysis (Marguerat et al., 2008; Marioni et al., 2008). Compared to microarray technology, RNA-seq does not depend on existing gene annotation for prior probe design, avoids the background noise from the hybridization of microarrays, and has a broader dynamic range of expression levels (Marguerat and Bähler, 2010; Metzker, 2010). As such, RNA-seq has shown strong potential to replace microarray technology for transcriptome analysis. RNA-seq directly sequences transcripts and obtains tens of millions of short reads from the transcript population of interest (Mortazavi et al., 2008). After mapping the sequenced reads to reference sequences, the mapped reads along each transcript are counted, allowing for the digital measurement of transcription levels to be estimated. Changes in isoform expression levels of alternatively spliced genes are of functional importance in particular biological processes. For example, there is much evidence that the development of many complex diseases is highly correlated with changes in isoform expression levels (Humbert et al., 2007; Beyer et al., 2008). Therefore, accurately estimating isoform expression levels remains an important challenge for understanding complicated biological mechanisms. In order to infer isoform expression levels, rSeq models the distribution of read counts for exons as a Poisson distribution with uniform sampling across each transcript (Jiang and Wong, 2009). A Poisson-Gamma hierarchical model has been proposed for multi-sample RNA-seq data to estimate isoform-specific expression. This model assumes that the reads are distributed uniformly across the genome, and hence reads are sampled independently and uniformly from every possible nucleotide in the sample (Vardhanabhuti et al., 2013). However, due to a number of biases, such as the 5'- and 3'-end biases, priming or GC bias, and so on, the uniform assumption of read distribution is untenable, and moreover, these various biases cause the non-uniform read distribution to be illogical, which causes the non-uniformity read distribution (Hansen et al., 2010; Zheng et al., 2011). This leads to difficulties in estimating isoform expression levels. Many methods are proposed that adopt different bias correction strategies to overcome these problems. For example, NURD adopts a global bias curve for all genes and local bias curves, which are estimated using non-parametric models (Wu et al., 2011). mseq predicts the variable rate for different positions based on local sequences (Li et al., 2010). POME incorporates the base-specific variation and between-base dependence that affect the read coverage profile throughout each transcript (Hu et al., 2012). WemIQ assigns different weights to reads from different gene regions when calculating the weighted log-likelihood (Zhang et al., 2015). All these aforementioned methods use the Poisson distribution to model the read counts and treat the bias values as weight factors to the Poisson rate. Another category of methods usually adopts generative statistic models. Cufflinks, one of the most commonly used tools, uses a variable length Markov model to learn sequence-specific biases on surrounding sequences, and calculates the positional bias according to the relative position of fragments in the transcript sequences (Trapnell et al., 2010). BitSeq uses the same bias correction strategy

as Cufflinks (Glaus et al., 2012). Both methods use the bias weight to select a fragment of a specific length for a given transcript. RSEM uses the empirical read start position distribution to represent the non-uniform read distribution, which depends on the fraction of length along the transcript in a given start position (Li and Dewey, 2011). Tigar2 considers the nucleotide character of the corresponding reference sequences (Nariai et al., 2014). These methods each use their own bias correction strategies, and improve the accuracy in estimating isoform expression levels. However, in real data, we find that the read distribution is highly correlated across samples (Suo et al., 2014). Figure 1 shows that the variation patterns of non-uniform read distributions are almost consistent across three biological samples. All of the preceding methods and corresponding bias correction strategies ignore the fact that the read distribution is highly consistent across multiple samples. In current RNA-seq experiments, multi-sample RNA-seq datasets are ubiquitous. It is now common for an RNA-seq experiment to sequence the whole transcriptomes of samples obtained from many technical and biological replicates. Variations in technical replicates arise from the different steps in RNA-seq experiments, such as library construction and amplification. Additionally, variations in biological replicates may be intrinsic to the biology (Ares Jr, 2014). More replicates can improve the biological or statistical significance. Thus, a combined treatment approach for multi-sample RNA-seq data may help to improve isoform expression level data more so than separate treatments, since the multiple samples can be correlated.

For processing multi-sample RNA-seq data, another limitation of current methods is that of dealing with each single sample separately. Due to the effects of noisy reads, the expression levels of the same isoform from multiple samples are usually biased, especially for isoforms that are lowly expressed. For example, for any given multi-sample RNA-seq dataset, several noisy reads may fall into a discriminative exon, which may result in an isoform that contains this exon to change from non-expressed to lowly expressed. However, there may be no reads for this exon in the rest of the samples, and the corresponding isoform is not expressed. It is common sense that the same isoform should have similar expression level across multiple samples. When combining multiple samples, we may be able to infer that this lowly expressed isoform is caused by noisy reads, and is not actually expressed (Roberts et al., 2011). Therefore, a combined treatment for multi-sample RNA-seq data can help to reduce the effect of noisy reads, and to identify the isoforms that are indeed lowly expressed. Furthermore, when an alternatively spliced gene that contains multiple isoforms is expressed, often only partial isoforms within the gene are expressed while the rest are not. The relationship between the expression level of a gene and its corresponding isoforms has sparse specificity. While methods for detecting novel isoforms take this prior biological information into consideration (Behr et al., 2013; Mezlini et al., 2013), most current isoform expression estimation methods do not take this into account.

In the current study, we propose a structured sparse regression method (SSRSeq) to estimate isoform expression levels. First, we design a bias correction strategy, which creates a multi-sample bias curve. This curve represents the general tendency of the non-uniform read distribution for all genes across multiple samples. Then, we add a structured sparse regularization, l_2, l_1 -norm, to the SSRSeq model to combine multiple samples. The l_2, l_1 -norm regularization not only reduces the effect of noisy reads, especially for lowly expressed genes and isoforms, but also incorporates the sparse specificity between the expression level of a gene and its corresponding isoforms. Finally, we use four real RNA-seq datasets to evaluate our approach and compared it with other start-of-the-art methods.

MATERIAL AND METHODS

Notations

Assume G is the set of genes. For gene $g \in G$, it has K isoforms and M exons with lengths l_1, \dots, l_m . The gene structure can be represented using a binary matrix $A = (a_{ik})_{M \times K}$. Each isoform corresponds to a column in this matrix, where $a_{ik} = 1$ or $a_{ik} = 0$ represents that the i -th exon is included or excluded in the k -th isoform, respectively. If an exon is contained in more than one isoform with different lengths, we split the exon into non-overlapping parts and treat each part as a separate exon.

To facilitate fast computations, we first summarize the number of reads falling into each exon. We assume that the RNA-seq dataset has N samples. We let y_{ij} represent the observed number of reads from the j -th sample that fall in the i -th exon. According to the principles of RNA-seq experiments and gene structure, y_{ij} is proportional to both the exon length l_i and the sum of the expression levels of all isoforms containing the i -th exon, where x_{kj} is the expected number of reads of the k -th isoforms in the j -th sample, and represents the expression level of the corresponding isoforms (Equation 1). After obtaining x_{kj} , we can measure the expression level as RPKM (Reads Per Kilobase of exon model per Million mapped reads). If the j -th sample has a total of w_j mapped reads, then the expression level (in RPKM) is $10^9 \times x_{kj} / w_j$.

$$y_{ij} = l_i \sum_{k=1}^K a_{ik} x_{kj} \quad (\text{Equation 1})$$

Multi-sample bias curve

For RNA-seq data, the read counts of each gene are highly non-uniformly distributed, but the variation patterns are almost consistent across multiple samples. Figure 1 shows the counts of a randomly selected gene ENSG0000089220 in three biological replicates from Human Brain dataset. It can be seen that the read distributions are highly non-uniform and highly consistent across the three biological replicates.

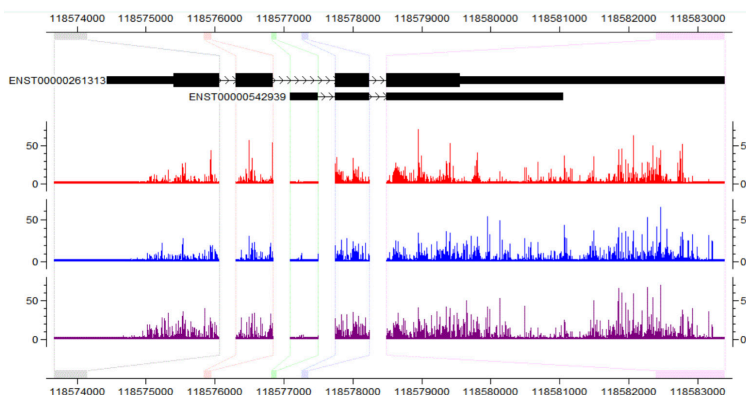


Figure 1. Read counts for each exonic nucleotide position in the CisGenome Browser (Ji et al., 2008) along gene ENSG0000089220 in three biological replicates. The gene ENSG0000089220 has two isoforms, ENST00000261313 and ENST00000542939, and the gene structure is from the Ensembl database (NCBI37/hg19).

We propose a multi-sample bias curve (MSBC) to characterize the read distribution along a gene across multiple samples. The MSBC represents the general tendency of the non-uniform read distribution for the whole transcriptome, and reflects the relative read distribution bias from the 5'- to 3'-ends of a gene. We separately deal with each single sample at the beginning of the MSBC calculation procedure. For each single sample, we use those genes that only contain a single isoform to estimate the MSBC. Since the multiple isoforms within a gene can affect the read distribution due to the gene structure, which does not reflect the general tendency of the read distribution along the gene, we filter out all multi-isoform genes according to gene annotation databases, including RefGene (Pruitt et al., 2012), Ensembl (Hubbard et al., 2002), and UCSC annotation (Karolchik et al., 2003). Because of the high uncertainty associated with lowly expressed genes, we exclude genes with too few (e.g., <50) of reads. In order to avoid the influence of local fluctuations in the read distribution, we uniformly split each gene into a small number of bins (e.g., 20), and summarize the read counts of each bin. Then, the read counts of each bin are normalized to a mean of 1 and averaged among the number of selected single isoform genes (Wu et al., 2011). Finally, we combine the multiple samples and use the polynomial regression model to fit the read counts of each bin. This gives a smooth MSBC, which represents the read distribution of a gene across multiple samples in the dataset. The curve reflects the read variation patterns of a gene with high consistency across multiple samples, of which the variation may be caused by the experiment protocol. The procedure for calculating the MSBC is shown in Figure 2.

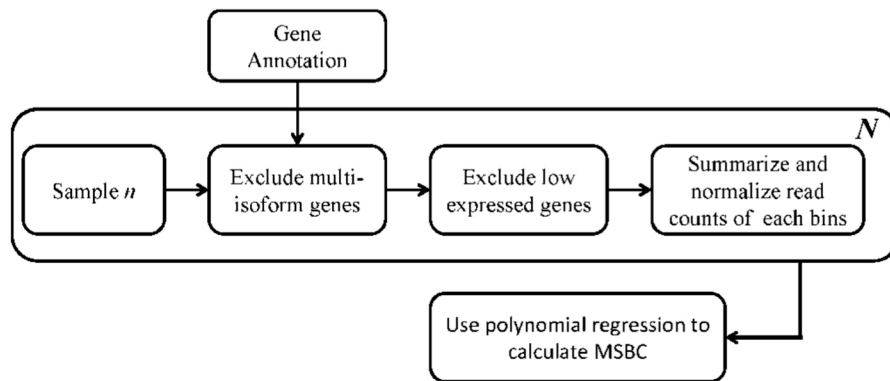


Figure 2. Procedure for calculating the MSBC.

An example of the MSBC curve obtained after using the polynomial regression model to fit the read counts of each bin across multiple samples is presented in Figure 3. From this bias curve, we can see that the 5'- and 3'-ends of genes have obvious position-specific biases. Hence, we use the MSBC to compensate for the read count variations in estimating isoform expression levels. For this purpose, we divide the curve into several segments corresponding to the length ratios of the exons along the gene. In Figure 3, the i -th segment corresponds to the i -th exon in the gene. We calculate the bias weights of the start and the end of the i -th segment, w_i and w_{i+1} , respectively. Then, the mean value of the i -th segment, $(w_i + w_{i+1}) / 2$, is calculated as the weighted factor for the i -th exon. In the following section, we will incorporate the bias weight into our model to estimate isoform expression.

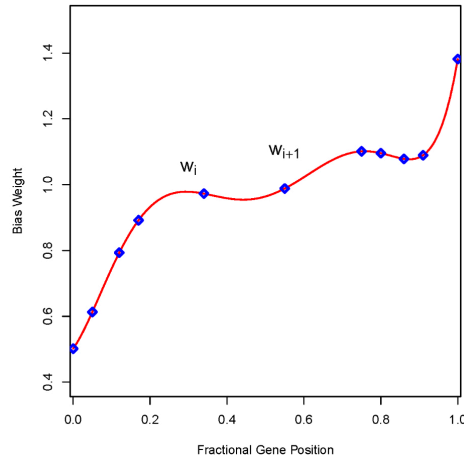


Figure 3. Illustration of the usage of MSBC. The solid line (red) is the MSBC obtained from the adult condition of the HB dataset. Each node represents the bias weight of the start or the end of a segment. For example, w_i and w_{i+1} represent the bias weight of the start and the end of the i -th segment, respectively.

SSRSeq model

Based on Equation 1, the expression calculation formula assumes uniform distribution of read counts along the reference sequence. However, the read counts of each gene are highly non-uniformly distributed. We propose the SSRSeq model by incorporating the bias weight, and thus rewrite Equation 1 as Equation 2.

$$y_{ij} = l_i \beta_i \sum_{k=1}^K a_{ik} x_{kj} \quad (\text{Equation 2})$$

where b_i is the bias weight of the i -th exon shared across multiple samples. The weight values are non-negative real numbers estimated from MSBC. Given all mapped reads and gene structure information of a gene, the expression levels $X = \{x_1, \dots, x_K\}$ of the isoforms can be estimated by minimizing the objective function shown in Equation 3, with the restrictions that $x_{kj} \geq 0$ for all $1 \leq k \leq K$ and $1 \leq j \leq N$. b_{ij} is an element of the bias weight matrix B and $b_{ij} = b_i a_{ij}$. Rather than the 0-1 indicator matrix A , the bias weight matrix B not only represents the structure information between a gene and its isoforms, but also contains the bias weight for the corresponding exons with non-zero elements a_{ij} .

$$\min_x \sum_{i=1}^M \sum_{j=1}^N \left(\frac{y_{ij}}{l_i} - \sum_{k=1}^K b_{ij} x_{kj} \right) \quad (\text{Equation 3})$$

Due to the effect of noisy reads in multi-sample RNA-seq data, the expression levels of the same isoform across multiple samples are usually biased, especially for lowly expressed isoforms. As mentioned above, the same isoform should have similar expression levels across multiple samples. Hence, a combined treatment for multi-sample RNA-seq data may help to reduce the effect of noisy reads, and accurately identify the true lowly expressed isoforms.

Furthermore, only certain isoforms within an alternatively spliced gene are expressed, whereas the rest are non-expressed isoforms. Therefore, the relationship between the expression level of a gene and its corresponding isoforms has sparse specificity. Considering this information, we add a structured sparse regularization, $l2, l1$ -norm, to the SSRSeq model to combine multiple samples (Meier et al., 2008), and rewrite the objective function as seen in Equation 4.

$$\min_x \sum_{i=1}^M \sum_{j=1}^N \left(\frac{y_{ij}}{l_i} - \sum_{k=1}^K b_{ij} x_{kj} \right) + \lambda \sum_{k=1}^K \left(\sum_{j=1}^N x_{kj}^2 \right)^{\frac{1}{2}} \quad (\text{Equation 4})$$

s. t. $x_{kj} \geq 0, 1 \leq i \leq M, 1 \leq j \leq N, 1 \leq k \leq K$

For convenient optimization, the objective function is equivalent to the matrix form seen in Equation 5.

$$\min_{X \in R^{K \times N}} \|D - BX\|_F^2 + \lambda \|X\|_{2,1} \quad (\text{Equation 5})$$

s. t. $x_{kj} \geq 0, D \in R^{M \times N}, B \in R^{M \times K}$

The $l2, l1$ -norm can make all variables of a row in matrix X be zero. This would mean that the corresponding isoform is non-expressed across multiple samples. The rest of the non-zero rows corresponds to expressed isoforms of the gene. Therefore, the $l2, l1$ -norm regularization not only reflects the sparse specificity between a gene and its corresponding isoforms, but also reduces the effect of noisy data for actual lowly expressed isoforms.

Due to $l2, l1$ -norm regularization, the objective function in Equation 5 is a non-smooth convex optimization problem. We adopt a popular optimization toolbox, SPAMS, for solving this intractable optimization problem (Mairal et al., 2010), and the solution \hat{X} is guaranteed to be a global optimal point.

The parameter λ controls the number of isoforms with non-zero expression levels in the solution. In the $l2, l1$ -norm constraint, a smaller value of λ will exert less restriction on the values of X , which results in more non-zero isoform expression levels. On the contrary, a larger value of λ prefers more zero rows, which represent corresponding non-expressed isoforms. In practice, a proper and fixed value of λ should be selected based on empirical data. In the current study, we set $\lambda = 1$ for all RNA-seq datasets.

Implementation

For RNA-seq data analysis, when reads are aligned to the reference transcriptome sequence, the process of expression estimation in SSRSeq can be divided into two stages, read alignment and expression estimation. The detailed workflow of SSRSeq is listed in Algorithm 1. SSRSeq has been implemented in Python and MATLAB. The alignment stage uses Python programming language, and the expression estimation stage is coded in MATLAB. SSRSeq makes use of parallel computing to improve the computation efficiency. The software and documentation are freely available at the website <http://parnec.nuaa.edu.cn/liux/ssreq.html>.

Algorithm 1

Algorithm 1: SSRSeq

Input: N alignment samples, Gene annotation

Read alignment:

Calculating MSBC using N alignment samples.
 Counting D for each gene in N alignment samples.
 Estimating b_i for each exon in each gene
 Expression estimation:
 repeat $\hat{X} = \operatorname{argmin}_x \| D - BX \|_2^2 + \lambda \| X \|_2, 1$
 until all genes are completed
 Output: The expression level of all isoforms in multiple samples.

RESULTS

Datasets

In the present study, four different RNA-Seq datasets were used to compare the performance of SSRSeq with three popular alternative methods, NURD (v.1.1.1), Cufflinks (v.2.2.1), RSEM (v.1.2.19), and one newly proposed method, Tigar2 (v.2.1). The four RNA-seq datasets can be freely downloaded from <http://sra.dnanexus.com/> with the accession numbers indicated below.

We use the well-studied Micro Array Quality Control (MAQC) project to validate the gene expression estimation from SSRSeq. MAQC project compares the multiple whole-genome gene expression across various commercial platforms (Shi et al., 2006), and is widely used to evaluate platform performance and to test various data processing methods (Canales et al., 2006; Li and Dewey, 2011). In the current study, we selected two RNA samples, the universal human reference (UHR) and the human brain reference (HBR). The Short Read Archie accession number was SRA010153 for the single-end (SE) dataset and SRA012427 for the paired-end (PE) dataset. MAQC project provided around 1000 genes validated by quantitative real-time polymerase chain reaction (qRT-PCR) experiments, which were regarded as the benchmark gene expression estimations that were obtained from diverse platforms and approaches. We also used the Ensembl annotation (NCBI37/hg19), and obtained 833 matching qRT-PCR validated genes.

A publicly available human breast cancer (HBC) dataset was used to validate the isoform expression estimation from SSRSeq and had the accession number SRA008403. This dataset contained two conditions, a human breast cancer cell line (MCF-7) and a norm cell line (HMR). Eight isoforms in five genes had been validated by qRT-PCR experiments for this dataset. We used the UCSC known Gene annotation (NCBI36/hg18) in order to obtain all annotation information for the eight qRT-PCR validated isoforms (Wang et al., 2008).

A mouse dataset was obtained from a study by Marioni et al. (2008) (SRA001030). The dataset contained data from three mouse tissues, including liver, skeletal muscle, and brain. Two technical replicates of the sample were sequenced on the Illumina/Solexa platform. We used the UCSC RefGene annotation (NCBI37/mm9), which contained 33608 isoforms for this dataset. Among the isoforms, 21.1% were from multi-isoform genes, and each gene had 1.39 isoforms on average.

Finally, the human brain (HB) dataset, downloaded from DDBJ (Kaminuma et al., 2011) with accession number SRA009447 was also used. This dataset contained two conditions, adult and fetal human brains, each of which had three biological replicates. Among the 163455 isoforms defined in the Ensembl annotation (NCBI37/hg19), 53.2% were from multi-isoform genes, and each gene had 4.69 isoforms on average.

Multi-sample bias curves describe SSRSeq

The MSBCs for each condition of the HB dataset are shown in Figure 4. The two MSBCs show strong distribution bias from the 5'- to 3'-ends of genes. We also observed that the two MSBCs had similar distribution patterns under the two conditions in the same experiment. Compared with the bias curve obtained from each single sample, the MSBC captured the common features from the multiple samples, especially at the 5'- and 3'-ends of genes.

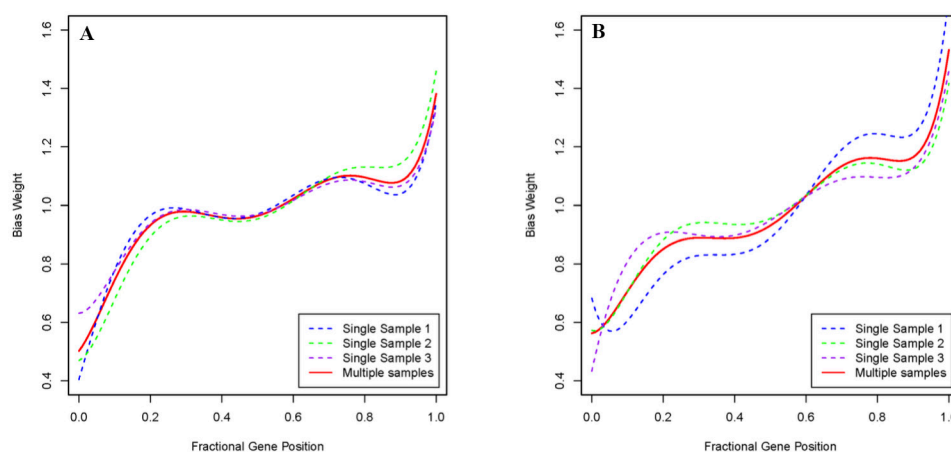


Figure 4. MSBCs obtained from the HB dataset for the (A) adult and (B) fetal conditions. The solid line (red) represents the MSBCs obtained from three samples under each condition. The dashed line represents the bias curve obtained from each single sample.

Robustness for noisy reads

Noisy reads may be caused by multiple factors including sequencing errors, mapping errors, and contamination by genomic DNA. SSRSeq adopts a structured sparse regularization to reduce the influence of noisy reads. Table 1 shows the isoform expression of gene ENSG00000166073, obtained from Tigar2 and SSRSeq. The expression of the first two isoforms from Tigar2, ENST00000485414 and ENST00000487571, are lowly expressed in samples 1 and 2, but are not expressed in the rest of samples. According to the biased expression across multiple samples, we think that the two isoforms are influenced by noisy reads, and are not actual lowly expressed isoforms. Cufflinks and RSEM resulted in similar expression profiles for the two isoforms. However, SSRSeq revealed that both isoforms are not expressed, and thus reduced the influence of noisy reads. Furthermore, SSRSeq also agrees with common sense that only partial isoforms within an alternatively spliced gene are expressed.

We further validated the ability of SSRSeq to reduce the influence of noisy reads at the gene expression level. We select gene ENSG00000037965, and estimated the gene expression using various methods. The results are shown in Table 2. From Table 2, we can see that except for SSRSeq, the other four methods resulted in low gene expression in some individual samples. We found that none or few reads were actually mapped to this gene. The mapped reads were most likely caused by noise, which resulted in biased gene expression. However, SSRSeq reduced this effect of noisy reads, and revealed that the gene was not expressed.

Table 1. Expression levels of the five isoforms within gene ENSG00000113811 estimated by Tigar2 and SSRSeq.

	Tigar2			SSRSeq		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
ENST00000485414	0	1.1656	0	0	0	0
ENST00000487571	0	0.2028	0.1438	0	0	0
ENST00000488746	20.0068	25.3979	26.1782	18.4631	25.1888	22.9183
ENST00000495461	3.1957	7.9404	6.8529	2.0734	5.0126	5.3042
ENST00000541726	0	2.6709	6.6144	4.3785	12.6145	13.8881

The data are from the adult condition in the HB dataset.

Table 2. Gene expression levels of ENSG00000037965 estimated by various methods.

Method	Sample 1	Sample 2	Sample 3
NURD	0.0111	0.0023	0.0212
Cufflinks	0	0.0666	0
RSEM	0.0312	0.0720	0.0201
Tigar2	0	0.0772	0.0107
SSRSeq	0	0	0

The data are from the adult condition in the HB dataset.

NURD, Cufflinks, RSEM, and Tigar2 estimated the expression for each individual RNA-seq sample and were easily influenced by noisy reads, while the method presented herein used a structured sparse regularization to combine multiple samples, which avoided this effect. These results demonstrate that SSRSeq is able to identify the true non-expressed or lowly expressed genes and isoforms.

Application on real RNA-seq data

In order to further evaluate the performance of SSRSeq in whole-genome expression profiles, we used four real datasets from different species and sequencing platforms. First, we used the mouse and HB datasets to show the consistency of isoform expression between multiple samples from various methods. All methods, except for SSRSeq, separately estimated isoform expression. We calculate R^2 correlation coefficients of the obtained isoform expressions between multiple samples as shown in Table 3. We observed that SSRSeq more consistently predicted isoform expression across multiple samples than the other alternative methods. We also noticed that the R^2 correlation coefficients from the mouse dataset were higher than those from the HB dataset. The reason for this may be that the multiple samples of the mouse dataset are technical replicates, while those of the HB dataset are biological replicates. As is well known, the variability in biological replicates is higher than that in technical replicates. Another reason may be that the number of isoforms in the HB dataset using Ensembl annotation is five times greater than that in the mouse dataset as more isoforms make it harder to accurately predict isoform expression.

Second, we used two datasets from the MAQC project to evaluate the gene expression calculations from various methods. The SRA010153 dataset contained two RNA samples, HBR and UHR. Each sample included seven technical replicates. The SRA012427 dataset contained three technical replicates from the UHR sample. Except for SSRSeq, the other methods separately estimated gene expression and averaged the results. Next, we calculated

the R^2 correlation coefficients between the average logarithm of the expression level of the 833 qRT-PCR validated genes and the qRT-PCR measurements. The results of the two datasets are presented in Table 4. We observed that SSRSeq outperformed the other four alternatives in the four different comparisons.

Table 3. R^2 correlation coefficients of estimated isoform expression levels by various methods between multiple samples in the mouse and HB datasets.

	NURD	Cufflinks	RSEM	Tigar2	SSRSeq
Mouse.Brain	0.942	0.943	0.953	0.949	0.974
Mouse.Liver	0.949	0.944	0.950	0.944	0.976
Mouse.Muscle	0.939	0.938	0.942	0.939	0.967
HB.Adult	0.749	0.753	0.751	0.737	0.784
HB.Fetal	0.747	0.751	0.756	0.742	0.772

The highest R^2 correlation coefficient for each comparison is in bold.

Table 4. R^2 correlation coefficients between the estimated gene expression levels by various methods and qRT-PCR measurements in the MAQC dataset.

	NURD	Cufflinks	RSEM	Tigar2	SSRSeq
SE.HBR	0.804	0.865	0.850	0.841	0.869
SE.UHR	0.815	0.872	0.871	0.864	0.878
PE.UHR	0.777	0.817	0.850	0.798	0.867

For each datasets, the abbreviations are as follows, SE: single-end and PE: paired-end. The highest R^2 correlation coefficient for each comparison is in bold.

Finally, we used the HBC dataset to evaluate the accuracy of the isoform expression estimation of our approach. Eight qRT-PCR validated isoforms related to five genes were deemed as benchmarks to compare the accuracy of various methods. The MCF-7 and HME cell lines contained seven and four technical replicates, respectively. We calculated R^2 correlation coefficients for each cell line and the results of all methods are shown in Table 5. Even though all methods resulted in relatively low correlation coefficients, SSRSeq presented the highest performance.

Table 5. R^2 correlation coefficients between the estimated isoform expression levels by various methods and the qRT-PCR measurements in the HBC dataset.

	NURD	Cufflinks	RSEM	Tigar2	SSRSeq
HME	0.119	0.529	0.493	0.507	0.616
MCF-7	0.193	0.249	0.211	0.210	0.411

The highest R^2 correlation coefficient for each comparison is in bold.

An example of the read distribution along the gene HIST1H2BD is shown in Figure 5. The read counts of each exon shown in Figure 5A and B are from two of the four samples in the MCF-7 condition of the HBC dataset. The proportions of the corresponding isoform expression levels estimated by various methods are listed in Table 6. Notably, SSRSeq produced the most consistent results with those of qRT-PCR experiments among the compared methods. The differences between SSRSeq and the other methods is clearly seen in Figure 5. Gene HIST1H2BD contained two isoforms, uc003ngr.1 and uc003ngs.1. The uc003ngr.1

isoform has a non-overlapping region in the 3'-end and the second exon of uc003ngr.1 is a discriminative exon, which are both useful to estimate isoform expression levels. Thus, the few reads mapped to the non-overlapping region of uc003ngr.1 are likely noisy reads, and resulted in the high proportion of isoform expression in the compared methods. However, SSRSeq reduced this effect, and accurately inferred the isoform expression level.

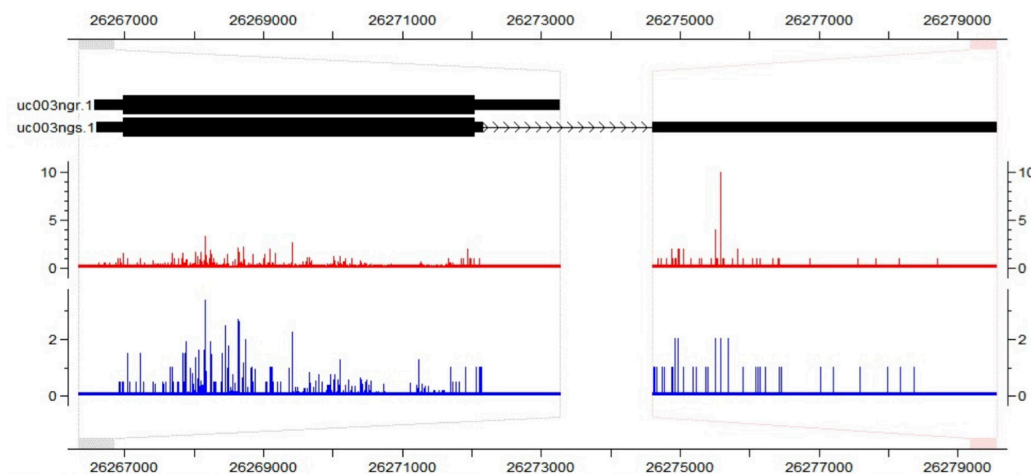


Figure 5. Read counts for each exonic nucleotide position in the CisGenome Browser along gene HIST1H2BD from the two samples in condition MCF-7 of the HBC dataset. The gene structure is from UCSC known Gene (NCBI36/hg18).

Table 6. Proportions of isoform expression levels of gene HIST1H2BD estimated by various methods in condition MCF-7 of the HBC dataset.

	qRT-PCR	NURD	Cufflinks	RSEM	Tigar2	SSRSeq
uc003ngr.1	1.6%	73.5%	66.8%	64.5%	63.3%	0%
uc003ngs.1	98.4%	26.5%	33.2%	35.5%	36.7%	100%

Model selection

SSRSeq uses the parameter λ to control the number of isoforms with non-zero expression levels in the solution. For the structured sparse regularization, l_2, l_1 -norm, a smaller value of λ can produce more non-zero isoform expression levels. On the contrary, a larger value of λ produces more non-expressed isoforms. As shown in Figure 6, as λ increases, the R^2 correlation coefficient between the estimated gene expression and the qRT-PCR experiment measurements decreases. However, the R^2 correlation coefficients of the gene/isoform expression between multiple samples increase, since most isoforms are forced to be non-expressed. Therefore, we need to balance these measurements. Therefore, according to Figure 6, $\lambda = 1$ is a proper and empirical choice.

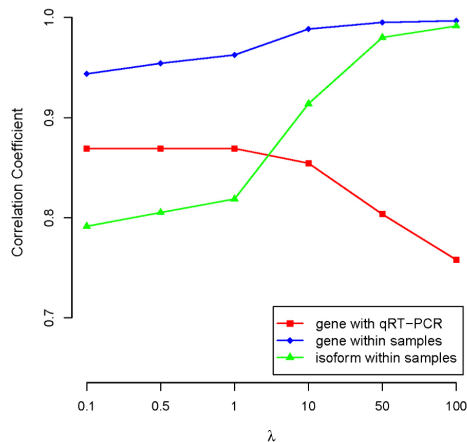


Figure 6. Results of expression estimations with the different λ selections in the HBR condition of the MAQC dataset. The red line represents the R^2 correlation coefficients between the estimated gene expression and the qRT-PCR measurements. The blue and green lines represent the R^2 correlation coefficients of the gene and isoform expression between multiple samples, respectively.

DISCUSSION

In the current study, we proposed the structured sparse regression method, SSRSeq, to estimate isoform expression levels using multi-sample RNA-seq data. We introduced bias curves across multiple samples, such as those of technical/biological replicates, and considered them as bias weights in SSRSeq. We used four real datasets to evaluate the performance of our method, and compared it with four other popular alternatives, NURD, Cufflinks, RSEM, and Tigar2. For real datasets, the proposed multi-sample bias curve represented the general tendency of non-uniform read distribution for all genes across multiple samples and presented a strong distribution bias from the 5'- to 3'-ends of genes. Moreover, SSRSeq used the structured sparse regularization, $l_{2,1}$ -norm, to reduce the effects of noisy reads and variance between multiple samples, especially for lowly expressed isoforms and genes. Concordance with qRT-PCR experiment measurements was also compared among the different methods, and revealed that our method produced more accurate isoform expression estimations.

In RNA-seq data analysis, detecting novel isoforms is an important and challenging task. Some approaches have been proposed to deal with this task for a given sample (Li et al., 2011a,b). However, this problem is often ill-posed since different combinations of isoforms may correctly explain the observed read counts, particularly at low coverage. If some isoforms are expressed in multiple samples potentially with different abundance, detecting novel isoform expression levels from multiple samples may result in better estimations. SSRSeq uses the structured sparse regularization, $l_{2,1}$ -norm, to obtain sparse specificity between the expression of a gene and its corresponding isoforms from multiple samples. Our model may easily be generalized to detect novel isoforms. Using splice junctions or genome annotations, we can enumerate all candidate isoforms, and then adopt the generalized SSRSeq model to detect the most possible isoforms. We will seek to verify this possibility in future research.

ACKNOWLEDGMENTS

Research supported by NSFC grants (#61170152), Jiangsu Provincial Qinglan Project, and the Fundamental Research Funds for the Central Universities (#CXZZ11 0217).

REFERENCES

- Ares Jr M (2014). Methods for processing high-throughput RNA sequencing data. *Cold Spring Harb. Protoc.* 2014: 1139-1148. <http://dx.doi.org/10.1101/pdb.top083352>
- Behr J, Kahles A, Zhong Y, Sreedharan VT, et al. (2013). MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics* 29: 2529-2538. <http://dx.doi.org/10.1093/bioinformatics/btt442>
- Beyer K, Domingo-Sábat M, Humbert J, Carrato C, et al. (2008). Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in Lewy body disease. *Neurogenetics* 9: 163-172. <http://dx.doi.org/10.1007/s10048-008-0124-6>
- Canales RD, Luo Y, Willey JC, Austermiller B, et al. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* 24: 1115-1122. <http://dx.doi.org/10.1038/nbt1236>
- Glaus P, Honkela A and Rattray M (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28: 1721-1728. <http://dx.doi.org/10.1093/bioinformatics/bts260>
- Hansen KD, Brenner SE and Dudoit S (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38: e131-e131. <http://dx.doi.org/10.1093/nar/gkq224>
- Hu M, Zhu Y, Taylor JM, Liu JS, et al. (2012). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics* 28: 63-68. <http://dx.doi.org/10.1093/bioinformatics/btr616>
- Hubbard T, Barker D, Birney E, Cameron G, et al. (2002). The Ensembl genome database project. *Nucleic Acids Res.* 30: 38-41. <http://dx.doi.org/10.1093/nar/30.1.38>
- Humbert J, Beyer K, Carrato C, Mate JL, et al. (2007). Parkin and synphilin-1 isoform expression changes in Lewy body diseases. *Neurobiol. Dis.* 26: 681-687. <http://dx.doi.org/10.1016/j.nbd.2007.03.007>
- Ji H, Jiang H, Ma W, Johnson DS, et al. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* 26: 1293-1300. <http://dx.doi.org/10.1038/nbt.1505>
- Jiang H and Wong WH (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25: 1026-1032. <http://dx.doi.org/10.1093/bioinformatics/btp113>
- Kaminuma E, Kosuge T, Kodama Y, Aono H, et al. (2011). DDBJ progress report. *Nucleic Acids Res.* 39: D22-D27. <http://dx.doi.org/10.1093/nar/gkq1041>
- Karolchik D, Baertsch R, Diekhans M, Furey TS, et al.; University of California Santa Cruz (2003). The UCSC genome browser database. *Nucleic Acids Res.* 31: 51-54. <http://dx.doi.org/10.1093/nar/gkg129>
- Li B and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323. <http://dx.doi.org/10.1186/1471-2105-12-323>
- Li J, Jiang H and Wong WH (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* 11: R50. <http://dx.doi.org/10.1186/gb-2010-11-5-r50>
- Li JJ, Jiang CR, Brown JB, Huang H, et al. (2011a). Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* 108: 19867-19872. <http://dx.doi.org/10.1073/pnas.1113972108>
- Li W, Feng J and Jiang T (2011b). IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* 18: 1693-1707. <http://dx.doi.org/10.1089/cmb.2011.0171>
- Mairal J, Jenatton R, Obozinski GR and Bach FR (2010). Network flow algorithms for structured sparsity. *Adv. Neural Inf. Process. Syst.* 23: 1558-1566.
- Marguerat S and Bähler J (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* 67: 569-579. <http://dx.doi.org/10.1007/s00018-009-0180-6>
- Marguerat S, Wilhelm BT and Bähler J (2008). Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.* 36: 1091-1096. <http://dx.doi.org/10.1042/BST0361091>
- Marioni JC, Mason CE, Mane SM, Stephens M, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18: 1509-1517. <http://dx.doi.org/10.1101/gr.079558.108>
- Meier L, Van De Geer S and Bühlmann P (2008). The group lasso for logistic regression. *J. Roy. Stat. Soc. Series B. Stat. Methodol.* 70: 53-71. <http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x>
- Metzker ML (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11: 31-46. <http://dx.doi.org/10.1038/nrg2626>

- Mezlini AM, Smith EJ, Fiume M, Buske O, et al. (2013). iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* 23: 519-529. <http://dx.doi.org/10.1101/gr.142232.112>
- Mortazavi A, Williams BA, McCue K, Schaeffer L, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628. <http://dx.doi.org/10.1038/nmeth.1226>
- Nariai N, Kojima K, Mimori T, Sato Y, et al. (2014). TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. *BMC Genomics* 15 (Suppl 10): S5. <http://dx.doi.org/10.1186/1471-2164-15-S10-S5>
- Pruitt KD, Tatusova T, Brown GR and Maglott DR (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 40: D130-D135. <http://dx.doi.org/10.1093/nar/gkr1079>
- Roberts A, Pimentel H, Trapnell C and Pachter L (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27: 2325-2329. <http://dx.doi.org/10.1093/bioinformatics/btr355>
- Shi L, Reid LH, Jones WD, Shippy R, et al.; MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24: 1151-1161. <http://dx.doi.org/10.1038/nbt1239>
- Suo C, Calza S, Salim A and Pawitan Y (2014). Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data. *Bioinformatics* 30: 506-513. <http://dx.doi.org/10.1093/bioinformatics/btt704>
- Trapnell C, Williams BA, Pertea G, Mortazavi A, et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511-515. <http://dx.doi.org/10.1038/nbt.1621>
- Vardhanabhuti S, Li M and Li H (2013). A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-Sample RNA-Seq Data. *Stat. Biosci.* 5: 119-137. <http://dx.doi.org/10.1007/s12561-011-9052-3>
- Wang ET, Sandberg R, Luo S, Khrebtkova I, et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470-476. <http://dx.doi.org/10.1038/nature07509>
- 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. <http://dx.doi.org/10.1093/bioinformatics/btq696>
- Zhang J, Kuo CC and Chen L (2015). WemIQ: an accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics* 31: 878-885. <http://dx.doi.org/10.1093/bioinformatics/btu757>
- Zheng W, Chung LM and Zhao H (2011). Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* 12: 290. <http://dx.doi.org/10.1186/1471-2105-12-290>