Bioinformatics, 33(19), 2017, 3018–3027
doi: 10.1093/bioinformatics/btx357
Advance Access Publication Date: 8 June 2017
Original Paper

OXFORD

Sequence analysis

DEIsoM: a hierarchical Bayesian model for identifying differentially expressed isoforms using biological replicates

Hao Peng[1,*,†], Yifan Yang[1,2,†], Shandian Zhe[1], Jian Wang[3], Michael Gribskov[1,2] and Yuan Qi[1,4]

[1]Department of Computer Science, [2]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA, [3]Eli Lilly and Company, Indianapolis, IN 46285, USA and [4]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

*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 Authors.
Associate Editor: Alfonso Valencia

Received on December 16, 2016; revised on May 5, 2017; editorial decision on May 29, 2017; accepted on June 2, 2017

Abstract

Motivation: High-throughput mRNA sequencing (RNA-Seq) is a powerful tool for quantifying gene expression. Identification of transcript isoforms that are differentially expressed in different conditions, such as in patients and healthy subjects, can provide insights into the molecular basis of diseases. Current transcript quantification approaches, however, do not take advantage of the shared information in the biological replicates, potentially decreasing sensitivity and accuracy.
Results: We present a novel hierarchical Bayesian model called Differentially Expressed Isoform detection from Multiple biological replicates (DEIsoM) for identifying differentially expressed (DE) isoforms from multiple biological replicates representing two conditions, e.g. multiple samples from healthy and diseased subjects. DEIsoM first estimates isoform expression within each condition by (1) capturing common patterns from sample replicates while allowing individual differences, and (2) modeling the uncertainty introduced by ambiguous read mapping in each replicate. Specifically, we introduce a Dirichlet prior distribution to capture the common expression pattern of replicates from the same condition, and treat the isoform expression of individual replicates as samples from this distribution. Ambiguous read mapping is modeled as a multinomial distribution, and ambiguous reads are assigned to the most probable isoform in each replicate. Additionally, DEIsoM couples an efficient variational inference and a post-analysis method to improve the accuracy and speed of identification of DE isoforms over alternative methods. Application of DEIsoM to an hepatocellular carcinoma (HCC) dataset identifies biologically relevant DE isoforms. The relevance of these genes/isoforms to HCC are supported by principal component analysis (PCA), read coverage visualization, and the biological literature.
Availability and implementation: The software is available at https://github.com/hao-peng/DEIsoM
Contact: pengh@alumni.purdue.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-seq is a powerful tool for investigating the transcriptomes of various organisms. There are many complex issues in RNA-seq and transcriptome analysis ranging from RNA-seq read correction (Le *et al.*, 2013), transcriptome assembly (Martin and Wang, 2011) to alternative splicing and gene fusion detection (Ozsolak and Milos, 2011). However, one of the most fundamental issues is to quantify and identify isoforms differentially expressed in two conditions, while each containing multiple replicates. Most DE isoform quantification methods treat each replicate independently, ignoring the fact that, because the underlying biological mechanism is the same in a given condition, the replicates tend to share similar expression patterns. DEIsoM improves DE isoform identification and quantification by catching the information shared between replicate samples; rather than separately estimating the isoform expression for each replicate, it captures the common expression pattern of the whole condition in one single model.

Although many computational tools have been developed for quantifying and identifying DE isoforms using RNA-seq data, nearly all approaches estimate the isoform abundance in each replicate separately, and do not attempt to actively capture the aforementioned shared information. For instance, Mixture of ISOforms (MISO) (Katz *et al.*, 2010) infers the isoform fractions for each replicate and evaluates the DE of every pair of replicates using the Bayes Factor, not considering replicates as a group. Additionally, MISO is slow due to its use of MCMC sampling, which is computationally challenging to adapt to the rapid growth in the amount of RNA-seq data (Kakaradov *et al.*, 2012). Dirichlet-Multinomial framework (DRIMSeq) (Nowicka and Robinson, 2016) infers the isoform fractions for each replicate in a Dirichlet-Multinomial model with a fixed hyperparameter and evaluates DE between two conditions by likelihood ratio test. Cufflinks (Trapnell *et al.*, 2012) quantifies the isoform abundance in individual replicates by maximum a posteriori (MAP) and detects DE isoforms by the hypothesis test based on Jensen-Shannon divergence. RNA-Seq by Expectation Maximization (RSEM) (Li and Dewey, 2011) estimates isoform abundance for each replicate using an Expectation Maximization (EM) algorithm. Empirical Bayesian Seq (EBSeq) (Leng *et al.*, 2013) then takes the expected counts from all replicates to fit a joint model and estimates the probability of DE for each isoform between multiple conditions. However, the variance of the expected counts stemming from ambiguous read mapping is simply lost in this process, compromising the DE isoform detection. Bayesian inference of transcripts from Sequencing data (BitSeq) (Glaus *et al.*, 2012; Hensman *et al.*, 2015) estimates the per condition mean isoform abundance from multiple replicates. However, BitSeq accomplishes this estimation in two stages rather than in an integrated model, which could potentially lose information when the "pseudo-data" from each fitted model in stage 1 is fed to the conjugate normal-gamma model in stage 2. Some other models do take the strategy of utilizing the shared information from multiple biological replicates, such as rMATS (Shen *et al.*, 2014), and MAJIQ (Vaquero-Garcia *et al.*, 2016). However, they are both exon-centric, quantifying and identifying alternative splicing at the exon level not the isoform level.

Here, we present DEIsoM, a hierarchical Bayesian model for quantifying and identifying DE isoforms between two conditions. Other than estimating the isoform abundance in each replicate separately, DEIsoM actively captures the shared information of perconditioned replicates in one principle framework. Specifically, DEIsoM uses a Dirichlet prior distribution to capture the shared information among replicates in each condition, and implements a fast Variational Bayesian (VB) method to gain computational efficiency instead of MCMC sampling when computing the posterior distributions of isoform fractions. Figure 1A shows a typical design for an RNA-Seq experiment with three replicates in each condition. Because we assume that the replicates in one condition share the same underlying biological mechanism, their expression patterns tend to be the same within a certain sample variance. We capture this common pattern through a Dirichlet prior with a tracable and effeciently updated hyperparameter. Additionally, we evaluate the DE isoforms by computing the Kullback–Leibler (KL) divergence between the posterior distributions of the two conditions, which is intrinsically fast in our model. Figure 1B gives a qualitative idea of how KL divergence is used to evaluate DE; the DE level is represented as the non-overlapping areas between the two posterior distributions.

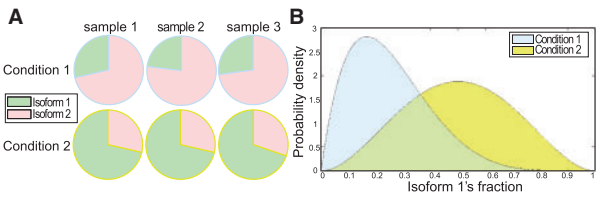

Fig. 1. DEIsoM estimation concept. (**A**) shows a typical RNA-Seq experimental setting targetted by DEIsoM. There are two conditions, each of which comprises three replicates shown as pie charts representing the expression fractions of two isoforms of a particular gene. We assume that the replicates in one condition are more likely to share a similar expression pattern, which will be captured by the Dirichlet prior distribution. (**B**) shows the posterior distribution of fractional isoform expression for each condition. The DE level of the isoform between two conditions can be represented by the non-overlapping regions (purely blue and yellow) under the two curves. In other words, the smaller the overlapping region is, the more distinct the two posteriors are, and the more differentially expressed the isoforms of this gene is. We measure this distinction by KL divergence, which is a widely recognized method to capture the difference between two probability distributions

Simulations in Section 3 demonstrate the superior performance of DEIsoM over alternative methods for quantifying and predicting DE isoforms, as well as the improved computational speed of VB method compared to MCMC sampling. Furthermore, on a real HCC dataset (Section 4), DEIsoM identifies HCC relevant DE isoforms which are supported by PCA, read coverage visualization, and the biological literature.

2 Materials and methods

DEIsoM consists of three parts: the hierarchical graphical model for isoform quantification (Section 2.1), the VB algorithm for model estimation (Section 2.2) and the identification of DE isoforms between two conditions (Section 2.3).

2.1 Model

Suppose we have collected RNA-seq data from M replicates in each condition. For the m^{th} replicate, there are in total $N^{(m)}$ paired-end reads that can be aligned to a given gene with K isoforms. Here, we utilize the previous annotated or assembled isoforms, so K is known for each gene. We use a K-dimensional binary vector, $\mathbf{R}_n^{(m)}$, to represent the read alignment to isoforms. If the n^{th} read from the m^{th} replicate maps to the k^{th} isoform, the k^{th} element of $\mathbf{R}_n^{(m)}$, $R_{n,k}^{(m)}$, is set to be 1, and 0 otherwise. The unsequenced fragment length between the n^{th} paired-end reads is denoted as $\boldsymbol{\lambda}_n^{(m)} = [\lambda_{n,1}^{(m)}, \ldots, \lambda_{n,K}^{(m)}]$.

First, we model how a read is generated from an isoform. We use a binary random variable $Z_{n,k}^{(m)}$ to represent whether the n^{th} read of the m^{th} replicate is actually generated from the k^{th} isoform. We call $Z_{n,k}^{(m)}$ the latent read origin. Although a read can map to multiple isoforms, it can only be sequenced from one isoform. Therefore, $\mathbf{Z}_n^{(m)}$ is a K-dimensional vector with exactly one element equal to 1 and all the others equal to 0, where $\sum_{k=1}^{K} Z_{n,k}^{(m)} = 1$. We assume that for the m^{th} replicate, $\mathbf{Z}_n^{(m)}$ follows a multinomial distribution $p(\mathbf{Z}_n^{(m)}|\boldsymbol{\psi}^{(m)})$, where $\boldsymbol{\psi}^{(m)}$ is a K-dimensional vector representing the fractions of isoforms in the m^{th} replicate for a given gene. Thus, $\psi_k^{(m)} \in [0,1]$ for all k and $\sum_{k=1}^{K} \psi_k^{(m)} = 1$. The fractions of isoforms $\{\boldsymbol{\psi}^{(m)}\}_{m=1..M}$ can vary among replicates, but we assume that the replicates all follow the same Dirichlet prior distribution $p(\boldsymbol{\psi}|\boldsymbol{\alpha})$ in each condition. Different from MISO, which uses one fixed prior $p(\boldsymbol{\psi})$ for each replicate, DEIsoM shares the same prior among replicates. The underlying reason is that the distributions of isoforms from different replicates of the same condition are not independent, but share some common patterns. DEIsoM summarizes the shared information in the hyperparameter $\boldsymbol{\alpha}$. In Section 2.2, we will further explain how the hyperparameter $\boldsymbol{\alpha}$ is updated using the information from all replicates.

We assume that the observed read alignments $R_{n,k}^{(m)}$ and the unsequenced fragment length $\lambda_{n,k}^{(m)}$ are conditionally independent given the corresponding latent read origin $\mathbf{Z}^{(m)}$ and some fixed parameters Θ:

$$p(R_{n,k}^{(m)}, \lambda_{n,k}^{(m)}|Z_{n,k}^{(m)}, \Theta) = p(R_{n,k}^{(m)}|Z_{n,k}^{(m)}, \Theta)p(\lambda_{n,k}^{(m)}|\Theta)$$

where Θ includes l_k, L, μ and σ^2. l_k is the length of the k^{th} isoform; L is the sequenced read length; μ and σ^2 are the mean and variance of $\lambda_n^{(m)}$ respectively. The first part, $p(R_{n,k}^{(m)}|Z_{n,k}^{(m)}, \Theta)$, represents the probability that a read can be aligned to a specific region of the k^{th} isoform conditioned on whether it is generated from this isoform. If the n^{th} read is generated from the k^{th} isoform, this read is assumed to be uniformly generated from one of all the possible positions in this isoform. Otherwise, $p(R_{n,k}^{(m)}|Z_{n,k}^{(m)}, \Theta)$ is 0. The number of all possible positions is $\tilde{l}_{n,k}^{(m)} = l_k - (2L + \lambda_{n,k}^{(m)}) + 1$, Then the conditional distribution is:

$$p(R_{n,k}^{(m)} = 1|Z_{n,k}^{(m)}, \Theta) = \begin{cases} 1/\tilde{l}_{n,k}^{(m)} & \text{if } Z_{n,k}^{(m)} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The second part, $p(\lambda_{n,k}^{(m)}|\Theta)$, is the probability of observing a paired-end read with unsequenced length $\lambda_n^{(m)}$, which follows a normal distribution with mean μ and variance σ^2. Both μ and σ^2 can be given or estimated from the aligned RNA-seq data. As a result, we have the following generative process for each of M replicates (Fig. 2):

1. $\boldsymbol{\psi}^{(m)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For each of $N^{(m)}$ reads:
   a. $\mathbf{Z}_n^{(m)} \sim \text{Multinomial}(1, \boldsymbol{\psi}^{(m)})$
   b. $R_{n,k}^{(m)} \sim p(R_{n,k}^{(m)}|Z_{n,k}^{(m)}, \Theta)$
   c. $\lambda_{n,k}^{(m)} \sim \text{Normal}(\mu, \sigma^2)$

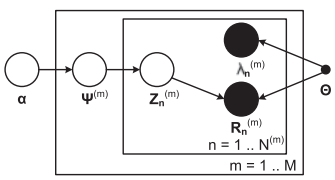

Fig. 2. The graphical model representation of DEIsoM

2.2 Estimation

To compute the posterior distribution of isoform fractions and read assignments,

$$p(\boldsymbol{\psi}, \mathbf{Z}|\mathbf{R}, \boldsymbol{\alpha}, \Theta) = \frac{p(\boldsymbol{\psi}, \mathbf{Z}, \mathbf{R}|\boldsymbol{\alpha}, \Theta)}{p(\mathbf{R}|\boldsymbol{\alpha}, \Theta)}$$

we need to compute the denominator:

$$p(\mathbf{R}|\boldsymbol{\alpha}, \Theta) = \prod_m \int p(\boldsymbol{\psi}^{(m)}|\boldsymbol{\alpha}) \prod_n \sum_k \Big[p(Z_{n,k}^{(m)} = 1|\psi_k^{(m)}) \times p(R_{n,k}^{(m)}, \lambda_{n,k}^{(m)}|Z_{n,k}^{(m)} = 1, \Theta)\Big] \mathrm{d}\boldsymbol{\psi}^{(m)}$$

which is computationally intractable, so we have to use approximate inference techniques, such as Markov Chain Monte Carlo (MCMC) sampling method or Variational Bayesian method. Classical MCMC methods may take a long time to converge due to the high correlation between the latent variables (Section 3.2). The Variational Bayesian method (Jordan *et al.*, 1999) tends to be faster and better scalable to large data for many graphical models. The VB algorithm approximates the intractable posterior p by a proposed distribution q, where q belongs to a family of distributions controlled by the variational parameters. We can optimize the variational parameters to minimize the Kullback-Leibler divergence between q and the posterior p, $\text{KL}(q||p)$. This is equivalent to maximizing a variational evidence lower bound. In such a way, the inference problem is cast to an optimization problem, which can be efficiently solved by gradient-based optimization algorithms.

For our model, we propose a family of variational distributions, which has the form:

$$q(\boldsymbol{\psi}, \mathbf{Z}) = \prod_m q\left(\boldsymbol{\psi}^{(m)}; \boldsymbol{\beta}^{(m)}\right) \prod_n q\left(\mathbf{Z}_n^{(m)}; \mathbf{r}_n^{(m)}\right),$$

where $q(\boldsymbol{\psi}^{(m)}; \boldsymbol{\beta}^{(m)})$ is a Dirichlet distribution parameterized by $\boldsymbol{\beta}^{(m)}$ and $q(\mathbf{Z}_n^{(m)}; \mathbf{r}_n^{(m)})$ is a multinomial distribution parameterized by $\mathbf{r}_n^{(m)}$.

We use the following iterative variational EM algorithm updates to find the optimal parameters for our model:

1. (E-step) For each replicate, estimate the variational parameters $\mathbf{r}_n^{(m)}$, $\boldsymbol{\beta}^{(m)}$;
2. (M-step) Maximize the variational evidence lower bound with respect to the hyperparameter $\boldsymbol{\alpha}$.

In E-step, we estimate the posterior distribution using a very commonly used algorithm, coordinate ascent variational inference (CAVI) (Bishop, 2006). We iteratively update:

$$r_{n,k}^{(m)} = \frac{\rho_{n,k}^{(m)}}{\sum_{l=1}^{K} \rho_{n,l}^{(m)}} \text{ and } \beta_k^{(m)} = \alpha_k + \sum_{n=1}^{N^{(m)}} r_{n,k}^{(m)} \tag{1}$$

where

$$\boldsymbol{\rho}_{n,k}^{(m)} = p\left(R_{n,k}^{(m)}, \lambda_{n,k}^{(m)}|Z_{n,k}^{(m)} = 1, \Theta\right) \times \exp\left[F\left(\beta_k^{(m)}\right) - F\left(\sum_{l=1}^{K} \beta_l^{(m)}\right)\right] \tag{2}$$

and F denotes the digamma function which is the derivative of the log-gamma function.

In M-step, we can use the Newton-Raphson method to update the hyperparameter $\boldsymbol{\alpha}$. This method is widely used for parameter estimation of models with Dirichlet priors (Blei *et al.*, 2003; Minka,

2000; Ronning, 1989). Here, we initialize the hyperparameter $\boldsymbol{\alpha} = 1$. The Newton–Raphson method finds the stationary point of an objective function using the iterative updates:

$$\boldsymbol{\alpha}_{\text{new}} = \boldsymbol{\alpha}_{\text{old}} - \boldsymbol{H}(\boldsymbol{\alpha}_{\text{old}})^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_{\text{old}}) \tag{3}$$

where $\boldsymbol{g}$ and $\boldsymbol{H}$ denote the gradient and the Hessian matrix of the objective function respectively. However, some new α_k may become non-positive during the iterative updates, which is invalid for Dirichlet distributions. Therefore, instead of working on $\boldsymbol{\alpha}$ directly, we update $\log(\boldsymbol{\alpha})$ first and then take the exponential of it. Let $\boldsymbol{\gamma} = \log(\boldsymbol{\alpha})$. The gradient and the Hessian of the variational lower bound with respect to $\boldsymbol{\gamma}$ can be computed as:

$$g_k(\boldsymbol{\gamma}) = M\left(F\left(\sum_{l=1}^{K}\alpha_l\right) - F(\alpha_k)\right)\alpha_k + \sum_{m=1}^{M}\alpha_k\left(F\left(\beta_k^{(m)}\right) - F\left(\sum_{l=1}^{K}\beta_l^{(m)}\right)\right) \tag{4}$$

$$H_{i,j}(\boldsymbol{\gamma}) = M\left(F'\left(\sum_{l=1}^{K}\alpha_l\right)\alpha_i\alpha_j\right) + \sigma(i,j)\Delta_i(\alpha) \tag{5}$$

where we define $\sigma(i,j) = 1$ if $i = j$, otherwise $\sigma(i,j) = 0$, F' is the trigamma function, and

$$\Delta_i(\boldsymbol{\alpha}) = M\left(F\left(\sum_{l=1}^{K}\alpha_l\right) - F'(\alpha_i)\alpha_i - F(\alpha_i)\right)\alpha_i + \alpha_i\sum_{m=1}^{M}\left(F\left(\beta_i^{(m)}\right) - F\left(\sum_{l=1}^{K}\boldsymbol{\beta}_l^{(m)}\right)\right) \tag{6}$$

A drawback of taking the logarithm is that we can no longer use the special structure of Hessian to compute $\boldsymbol{H}^{-1}\boldsymbol{g}$ efficiently as in Blei *et al.* (2003). Since Hessian computation can be expensive for large K, we update $\boldsymbol{\gamma}$ with L-BFGS method using the gradient only. Updates for $\boldsymbol{\alpha}$ will terminate when the maximum number of iterations is reached or the change in evidence lower bound is smaller than our threshold.

2.3 Identification

The DE level of an isoform can be represented as the difference between the posterior distributions of isoform fractions under two conditions. As used in the Variational Bayesian method, KL divergence measures the difference between any two distributions. Therefore, we compute the KL divergence between the posterior distributions of isoform fractions under the two conditions to evaluate the DE level of the isoforms. A higher KL divergence implies that the isoforms of this gene are more differentially expressed under the two conditions. Specifically, we train the model and estimate the posterior distribution $p(\boldsymbol{\psi}|\mathbf{R}, \boldsymbol{\alpha}, \Theta)$ with data from healthy and diseased conditions respectively. As described in Section 2.2, although the exact posterior distribution cannot be computed, we use the approximate posterior distributions from two conditions, $q(\boldsymbol{\psi}; \boldsymbol{\beta})$ and $q'(\boldsymbol{\psi}'; \boldsymbol{\beta}')$, to compute the KL divergence. Because $q(\boldsymbol{\psi}^m)$ or $q'(\boldsymbol{\psi}^m)$ are independent Dirichlet distributions, the KL divergence, D_{KL} can be computed analytically as:

$$D_{KL}(q||q') = \sum_{m=1}^{M}\left\{\log\frac{\sum_{k=1}^{K}\boldsymbol{\beta}_k^{(m)}}{\sum_{k=1}^{K}\boldsymbol{\beta}\prime_k^{(m)}} + \sum_{k=1}^{K}\log\frac{\Gamma\left(\boldsymbol{\beta}\prime_k^{(m)}\right)}{\Gamma\left(\boldsymbol{\beta}_k^{(m)}\right)} + \sum_{k=1}^{K}\left[\beta_k^{(m)} - \beta'^{(m)}_k\right]\left[F\left(\beta_k^{(m)}\right) - F\left(\sum_{l=1}^{K}\beta_l^{(m)}\right)\right]\right\} \tag{7}$$

To remove the asymmetry of D_{KL} between two conditions, we further compute the Jensen-Shannon divergence $D_{JS} = \frac{1}{2}[D_{KL}(q||q') + D_{KL}(q'||q)]$.

3 Simulations

In this section, we present four simulation studies to test that (1) whether DEIsoM benefits from the shared information from the multiple biological replicates compared with alternative methods; (2) whether the VB inference speeds up the computation without loss of accuracy; (3) whether DEIsoM is robust to different simulation settings; (4) whether the quantification of DEIsoM outperforms alternative methods under a more realistic setting.

3.1 Comparison of five methods on synthetic data

To test whether the shared information contributes to DE isoform detection, we generate synthetic data and compare DEIsoM with four commonly used programs: Cufflinks (v2.2.1), MISO (v0.5.3), RSEM (v1.2.30), and BitSeqVB (v0.7.5). The synthetic data are generated as follows. We first randomly select 200 genes (1395 isoforms) from the annotation of chromosome 1 in the hg19 human reference genome, in which 100 genes are labeled as containing DE isoforms and the rest are non-DE. We sample the expression levels of genes from a log-normal distribution (Gierliński *et al.*, 2015). Isoform fractions are generated from a symmetric Dirichlet distribution with $\boldsymbol{\alpha} = 1$, which means the chance of sampling any fraction of isoforms is equally probable. For instance, if there are three isoforms, the probability of sampling the isoform fraction as (0.1, 0.2, 0.7) is the same as (0.2, 0.3, 0.5). For DE isoforms, we draw two different samples for two conditions respectively; for non-DE, we draw only one sample shared by both conditions. To model the variation among replicates, we add Gaussian noise with a standard deviation equal to 10% of the expression level of each replicate. According to Standards, Guidelines and Best Practices for RNA-Seq V1.0[1], the number of paired-end RNA-Seq reads used in current studies is around 30 million per replicate. And for each tissue, it is generally expected more than 10, 000 genes are expressed (Consortium, 2015). Following the above empirical read numbers, we generate 600, 000 RNA-Seq reads for 200 genes using RNASeqReadSimulator [2] for each of five replicates in both conditions, using default settings.[3] To test the robustness of DEIsoM, we repeat the above simulation process 10 times. For RSEM, BitSeq, MISO, and DEIsoM, the simulated reads are mapped back to the reference transcriptom using Bowtie2 (Langmead and Salzberg, 2012). For Cuffdiff, the reads are mapped back to the hg19 reference genome using Tophat (Trapnell *et al.*, 2009). The machine used to run all experiments has two 8-Core Intel Xeon-E5 processors and 64 GB memory.

First, we compare the quantification performance of DEIsoM with MISO, Cuffdiff, RSEM and BitSeqVB in terms of the

1 Standards, Guidelines, and Best Practices for RNA-Seq V1.0 can be found at: https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

2 RNASeqReadSimulator is available at: http://alumni.cs.ucr.edu/liw/rnaseqreadsimulator.html

3 Our simulation code is available at: https://github.com/hao-peng/DEIsoM/tree/master/simulation

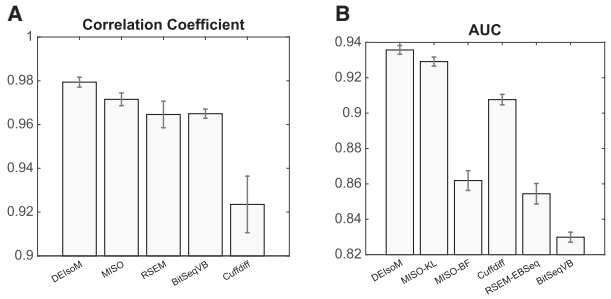

Fig. 3. RNA-Seq simulation studies. (**A**) Means and standard errors of correlation coefficients between the estimation and the ground truth in 10 replicates, using DEIsoM, MISO, RSEM, BitSeqVB and Cuffdiff. (**B**) Means and standard errors of AUCs of 10 repeated simulations for DEIsoM, MISO-KL, MISO-BF, Cuffdiff, RSEM-EBSeq, and BitSeqVB

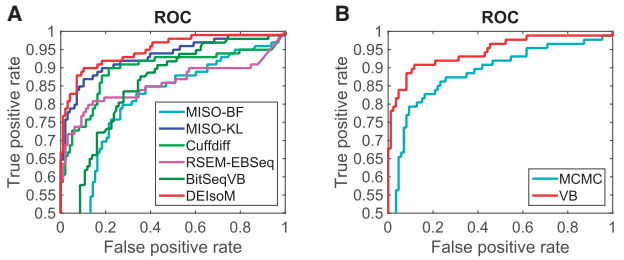

Fig. 4. RNA-seq simulation studies. (**A**) ROC curve comparison of MISO, Cuffdiff, RSEM-EBSeq, BitSeqVB, and DEIsoM from one run of 10 repeated experiments. For MISO we use two evaluation methods, MISO-KL and MISO-BF. MISO-KL denotes the average of KL divergences of the posteriors of isoform fractions. MISO-BF denotes the average of Bayes factors. (**B**) ROC curve comparison of VB and MCMC implementations of DEIsoM on the same dataset

correlations between the predicted isoform fractions and the ground truth on the synthetic data. Figure 3A summarizes the means and the standard errors of the correlation coefficients in 10 replicates. They show that the correlation coefficients in DEIsoM is higher than the alternative methods.

Second, we compare the DE isoform identification performance of DEIsoM with MISO, Cuffdiff, RSEM-EBSeq, and BitSeqVB in terms of the area under curve (AUC) of receiver operating characteristic (ROC) curves on the synthetic data. The ROC curves are computed based on different ranking criteria for the five methods. DEIsoM uses the KL divergence; MISO uses both the average of Bayes factors of all pairs of subjects (MISO-BF) and the average of KL divergences of posteriors of isoform factions (MISO-KL); Cuffdiff uses a log-fold-change based *P*-value; RSEM-EBSeq uses the Posterior Probability of Differential Expression (PPDE); BitSeqVB uses the Probability of Positive Log Ratio (PPLR). To make different criteria comparable, we take the minimum of all isoform *P*-values of the gene, and the maximum of all isoform PPDEs and PPLRs of the gene as the gene DE level. And we choose the "isoform-centric" mode for MISO. Also, PPLR is more sensitive to the upregulated DE isoforms than the downregulated ones by definition. Figure 4A shows the ROC curves for one of the 10 repeated experiments. Figure 3B summarizes the means and standard errors of the AUCs over 10 runs. They show that DEIsoM consistently outperforms MISO-BF, MISO-KL, Cuffdiff, and RSEM-EBSeq on the synthetic data under our settings.

Third, we compare the CPU time of DEIsoM, Cuffdiff, RSEM-EBSeq and BitSeqVB. The time we count is from the point we give the alignment files as input to the point that the programs generate the quantification results. We summarize it in Supplementary Table S1 for one run of the simulated data and the real data which will be discussed in Section 4. The numbers of hours used by the three algorithms are comparable, where Cuffdiff is always the fastest in all methods. However, DEIsoM has better DE isoform identification and quantification performance than the alternative methods, which is shown in both Section 3 and Section 4.

3.2 Comparison of VB and MCMC on synthetic data

To test whether the VB inference algorithm speeds up the computation over MCMC sampling without loss of accuracy, we compare the ROC curves and running time of the two implementations. We set the maximum iteration number as 1500 for both VB and MCMC. The burn-in time of MCMC is 150 iterations. Note that the MCMC sampling here is not completely the same as MISO. MISO combines the Metropolis-Hasting algorithm with a Gibbs sampler. We follow the same approach to estimate ψ, but we iteratively sample α from its posterior distribution given a non-informative prior which depends on all five replicates. Details of our MCMC sampling method are described in the Supplementary. The VB inference shows an advantage over MCMC in both the ROC curve and computing time within the limited number of iterations. Figure 4B shows the ROC curves for both implementations; VB inference achieves an AUC = 0.9445 in 1.4 CPU h, whereas the MCMC method has AUC = 0.8844 in 56 CPU h. Although MCMC theoretically can give samples from the exact target posterior distribution, it converges slowly on this dataset, which may cause inaccurate predictions and long running time. However, VB usually converges before the limit is reached under the same number of maximum iterations. Therefore, the VB method achieves a faster speed and a higher accuracy than the MCMC sampling.

3.3 Comparison of sensitivity of five methods

To demonstrate the robustness of DEIsoM, we vary the parameter of Dirichlet distribution α used for generating isoform fractions. When we increase α, the variance of generated isoform fractions under two conditions becomes smaller, but the mean remains the same. As a result, the difficulty of distinguishing DE genes from non-DE genes increases. In this experiment, we set $\alpha = 1$, **3**, and **5** and keep the other settings unchanged to simulate the data. We test all above five methods on the simulated reads to see whether they are sensitive to the change of α. Table 1 shows that as α increases, the AUCs of all methods decrease, since the task becomes harder. However, DEIsoM consistently outperforms the alternative methods throughout all α settings.

3.4 Comparison of abundance estimation

To test the quantification performance of DEIsoM under a more realistic setting, we simulate RNA-Seq reads using real data. Two RNA-Seq datasets of human stomach tissue were chosen from the ENCODE project[4]. Following the same percedure in Hensman *et al.* (2015), we estimate the abundance of 196, 317 transcripts using four models, RSEM, Cuffdiff, BitSeqVB and DEIsoM, as the ground truth for each scenario. By feeding the ground truth to Spanki (Sturgill *et al.*, 2013), we generate about 10 millions paired-end reads for each of the five replicates under each scenario. Four

4 The datasets from ENCODE project can be found at: https://www.encodeproject.org/experiments/ENCSR853WOM/ and https://www.encodeproject.org/experiments/ENCSR752UNJ/

Table 1. AUCs for MISO, Cuffdiff, RSEM-EBSeq, BitSeqVB, and DEIsoM on simulated data with different α

α	1	3	5
MISO-BF	0.849	0.727	0.673
MISO-KL	0.912	0.878	0.844
Cuffdiff	0.890	0.834	0.815
RSEM-EBSeq	0.873	0.798	0.762
BitSeqVB	0.807	0.771	0.704
DEIsoM	0.931	0.915	0.887

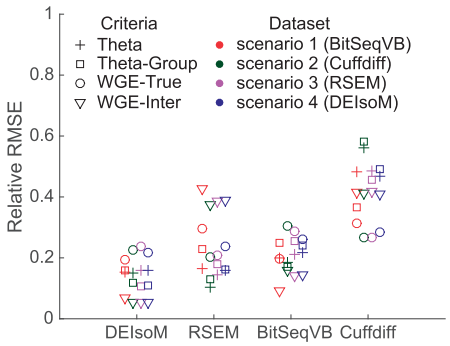

Fig. 5. Relative root mean squared errors of DEIsoM, RSEM, BitSeqVB and Cuffdiff on four simulated datasets. Theta: estimated transcript fractional expression compared with the ground truth for all the replicates. Theta-Group: mean estimated transcript fractional expression of the whole group compared with the true group mean. WGE-True: within-gene relative estimates compared with the ground truth. WGE-Inter: inter-replicate consistency of within-gene relative estimates

different evaluation criteria are used, see Supplementary S.2.1: Theta, Theta-Group, WGE-True and WGE-Inter. Theta measures the accuracy of transcript fraction estimation for all the replicates; Theta-Group measures the accuracy of transcript fraction estimation for the whole group; WGE-True measures the accuracy of within-gene relative fractional estimation; WGE-Inter measures the predictive consistency among all replicates. Figure 5 summarizes the relative root mean square errors (RMSE) of DEIsoM, RSEM, BitSeqVB, and Cuffdiff on four simulated datasets. They show that the DEIsoM RMSEs in both Theta-Group and WGE-Inter are lower than the other three methods, indicating that DEIsoM tends to give more consistent and accurate estimates for the whole condition. This result is consistent with the one in (Hensman *et al.*, 2015). A similar result evaluated by the relative mean absolute errors (MAE) is shown in Supplementary Figure S1.

4 Real data experiments and results

In this section, we test whether DEIsoM successfully identifies DE isoforms in real data. We apply DEIsoM and alternative programs to a hepatocellular carcinoma (HCC) RNA-seq dataset, and evaluate the predicted DE isoforms by PCA, read coverage visualization, and comparison to the biological literature. Aberrant alternative splicing is known to be involved in HCC (Berasain *et al.*, 2010), so DE isoforms should be present.

4.1 Data pre-processing

RNA-seq data was collected from nine pairs of HCC tumors and their matched adjacent normal tissues (Kan *et al.*, 2013; Sung *et al.*, 2012). The mRNA of each sample was extracted, amplified and sequenced. 150 base paired-end reads were generated and aligned to the hg19 human reference genome using RUM (RNA-Seq Unified Mapper) (Grant *et al.*, 2011). The aligned reads are used as input to three methods, Cuffdiff, RSEM-EBSeq, BitSeqVB, and DEIsoM, for DE isoform detection. MISO is not included because it cannot perform a group-wise analysis.

4.2 PCA

Because there is no exact ground truth for the HCC real data, we evaluate the quantification ability of each method by PCA plots. We first choose 38 significantly DE genes that are verified by polymerase chain reaction (PCR) from the previous publications (Dong *et al.*, 2009; Huang *et al.*, 2017; Wang *et al.*, 2015, 2017). For each gene, we sum up the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) of all the child isoforms as the gene expression. If the gene/isoform expressions associated with the HCC are correctly estimated, these gene/isoforms can be used as features to distinguish between the normal and tumor samples in PCA plots. Figure 6 shows that DEIsoM and RSEM can linearly separate tumor samples from their matched normal samples; BitSeqVB has one tumor sample (9) very closed to the normal cluster; Cuffdiff misses three tumor samples (4, 5, 6) in the normal cluster.

4.3 Read coverage visualization

To understand the expression patterns of the DE isoforms selected by DEIsoM, we visualize the read coverage on the hg19 reference genome. Because it may be possible to align a read to multiple isoforms, it is hard to determine the exact expression level of each isoform from the read coverage visualization. But it is possible to tell the change in isoform expression in some cases. A previous study successfully identified the genes with DE isoforms by testing the difference in read coverage between two conditions (Stegle *et al.*, 2010). Following the same logic, we assume that if the read coverage of a gene is similar in the two conditions, the isoforms of that gene will be predicted as non-DE. Otherwise, they are more likely to be DE.

First, we examine the read coverage of IGF2, a gene identified by DEIsoM as having DE isoforms. IGF2 is the 2^{nd} most DE gene identified by DEIsoM. Eight isoforms of IGF2 have been observed according to the human transcriptome annotation. Figure 7A and B show the read coverage of IGF2 in nine pairs of normal and tumor samples. Note that the reads aligned to the last two exons (in the box) can only contribute to isoform 4 (ENST00000300632). Figure 7B shows that the absolute numbers of reads aligned to the last two exons in all tumor samples are much lower than that in normal samples. Figure 7C is the same as Figure 7B but with an automatically scaled y-axis. (C) shows that in eight of nine tumor samples (1T, 2T, 4T – 9T), the fractions of reads aligned to the last two exons are much lower in the HCC samples than that in the normal samples. This indicates that IGF2 isoform 4 is down-regulated in HCC tumors. However, in the Cuffdiff results, this isoform has a *P*-value of 0.039 with rank 95; in RSEM-EBSeq, the PPDE equal to 1 out of 1147 DE isoforms all with PPDE = 1. But if we further rank by transcript real fold change (condition 1 over condition 2) as recommended, it ranks 671 out of 1147 DE isoforms.

Second, we show the read coverage of IGF2BP1, a gene identified by Cuffdiff as having DE isoforms. Isoform 1 (ENST00000290341) of IGF2BP1 is the 6^{th} most DE gene. Supplementary Figure S1 shows the read coverage of IGF2BP1 in normal and tumor samples. Note that the reads aligned to the last exon only contribute to isoform 1 (the box indicates the last exon).

However, only four of nine tumor samples show moderate differential expression of isoform 1 (lower than 500), and the expression level is near zero in all normal samples and five of the tumor samples (1T – 4T, 8T). Cuffdiff evaluates DE level using the log-fold-change between the conditions. This "fold" will be extremely large when the expression of one condition is near zero and the other is slightly higher. However, due to the low count numbers in both conditions, the confidence of calling this gene as having DE isoforms is low. Often, an empirical value is set to avoid low signals (NOTEST or LOWDATA). On the contrary, DEIsoM ranks IGF2BP1 as 244. Because both large sample variance and low read coverage lead to relatively "flat" posterior distributions in both normal and tumor conditions, which are close to the prior distribution. Thus, the KL divergence between two posterior distributions is small and the isoforms are not identified as DE.

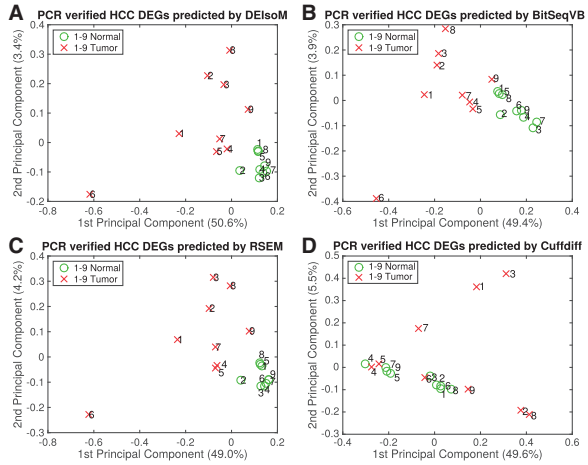

Fig. 6. PCA plots for nine pairs of HCC samples and their matched normal samples. Each sample is represented by a vector with 38 gene expressions. All these 38 genes are PCR verified DE genes in HCC. A, B, C, D are PCA plots using the estimations from DEIsoM, BitSeqVB, RSEM and Cuffdiff, respectively. Circle: normal sample. Cross: tumor sample. Percentage: the proportion of variance of the corresponding principle component

Lastly, we visualize the five least differentially expressed isoforms identified by DEIsoM, showing that the low ranked isoforms have very similar read coverage patterns in both normal and tumor samples. Supplementary Figure S2 shows COX16 has a similar read coverage pattern among all samples in both normal and tumor conditions. This is because a low KL divergence requires a high similarity between two posterior distributions of isoform fraction.

4.4 Biological relevance of predicted DE isoforms

To further understand the functions of DE isoforms selected by DEIsoM, we examine whether they are supported by HCC relevant literature. PubMed searches were performed using the keywords 'gene name + hepatocellular carcinoma'. Since most current experimental work focuses on the expression levels of genes rather than isoforms, we associate the DE isoforms identified by DEIsoM, Cuffdiff, RSEM-EBSeq and BitSeqVB with their gene names. Also, we assume that if the expression of a gene changes, it is very likely caused by a change of its isoforms. DE isoforms/genes are then categorized into four groups (3, 2, 1, 0) according to their relevance to HCC. 'Category 3' refers to a gene whose function in HCC has been well studied and can be used as a potential biomarker for prognosis or diagnosis. 'Category 2' indicates that differential expression of a gene has been detected *in vivo*, but not used as a biomarker. 'Category 1' indicates a gene whose function has only been studied *in vitro* but not in patient biopsies. 'Category 0' indicates a gene for which we found no HCC relevant literature.

First, we compare the number of genes that are HCC biomarkers (Category 3) in the predictions by DEIsoM, BitSeqVB, RSEM-EBSeq and Cuffdiff (the first four columns in Fig. 8). In the top 10 lists, more genes are identified as HCC biomarkers by DEIsoM than BitSeqVB, RSEM-EBSeq or Cuffdiff. Specifically, 6/10 genes identified by DEIsoM (ASS1, TTR, IGF2, AHSG, GPC3, CRP) vs. 4/10 genes identified by BitSeqVB (GPC3, AFP, IGF2BP3, UBE2C), 3/10 genes identified by Cuffdiff (SKP2, C-FOS, SOCS2) and 3/10 genes identified by RSEM-EBSeq (PEG10, TERT, ACAN) belong to Category 3.

Second, we have examined the six specific HCC biomarkers status (ASS1, TTR, IGF2, AHSG, GPC3, CRP) in the top 10 list of DEIsoM. Specifically, ASS1 is detected to be down-regulated in HCC liver

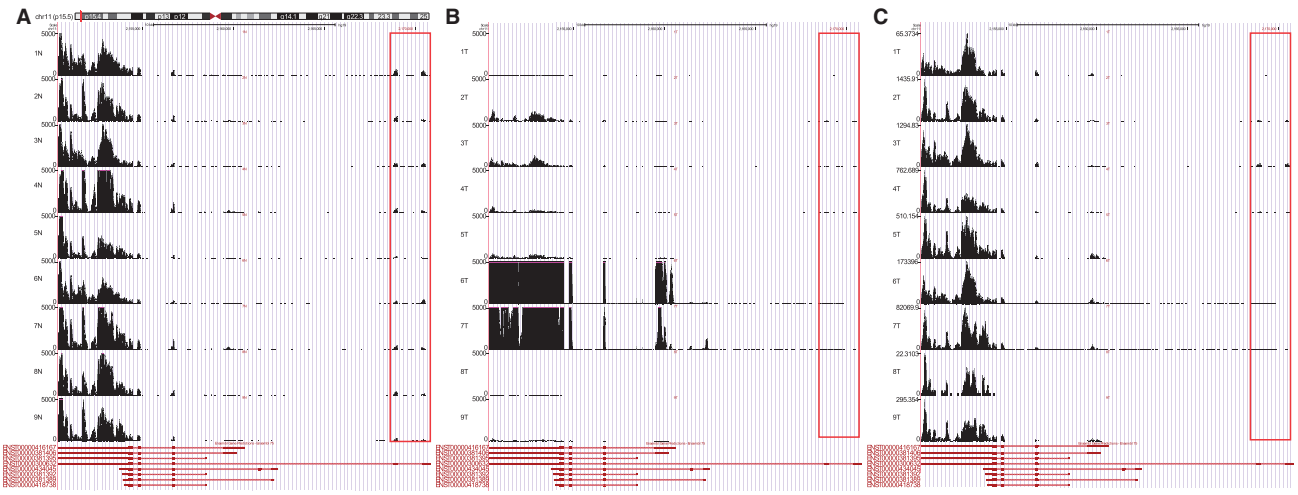

Fig. 7. Read coverage of IGF2 – a top selection by DEIsoM. The data was normalized across replicates by scaling the total number of reads to that of 1N (replicate 1 under normal condition). (**A**) Read coverage patterns of nine normal samples with *y*-axis scaled to 5000. (**B**) Read coverage patterns of nine matched tumor samples with *y*-axis scaled to 5000. (**C**) is the same as (B) but uses an automatically scaled *y*-axis. This illustrates that 1–5 and 8–9 tumor samples have very low read abundance in the last two exons, and the low signals are not due to the imposition of a fixed large *y*-axis scale. The exon positions of eight isoforms are listed under each panel

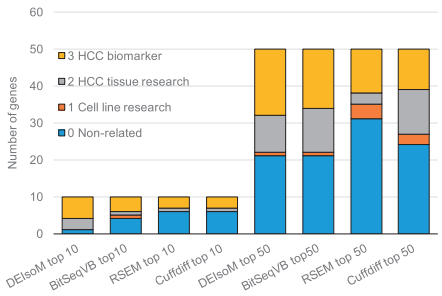

Fig. 8. HCC relevance of DE isoforms identified by DEIsoM, BitSeqVB, Cuffdiff, and RSEM-EBSeq. Relevance is defined as Category 3: HCC biomarkers, Category 2: DE genes verified in HCC tissues, Category 1: DE genes verified in HCC cell lines, and Category 0: HCC non-related genes. We analyze both the top 10 and top 50 selections for all four methods

samples, which can be used to predict metastatic relapse with a high sensitivity and specificity (Tan *et al.*, 2014); TTR is down-regulated in HCC patient serum (Qiu *et al.*, 2008); (Yim and Chung, 2010) state that both IGF2 and GPC3 are effective biomarkers for HCC—particularly, circulating IGF2 mRNA is positive in 34% of HCC patients and 100% correlated with the extrahepatic metastasis; GPC3 has been reported to interact with the Wnt signaling pathway to stimulate cell growth in HCC; GPC3 has also been used combined with PEG10, MDK, SERPINI1, and QP-C as a classifier that successfully distinguishes noncancerous hepatic tissues from HCCs (Yim and Chung, 2010); AHSG combined with two other HCC-associated antigens—KRT23 and FTL—can be used to diagnose HCC with sensitivity up to 98.2% in joint tests and specificity up to 90.0% in serial tests. (Wang *et al.*, 2009); CRP, an inflammatory cytokine, is highly expressed in HCC and its expression is correlated with tumor size, Child-Pugh function and survival time (Jang *et al.*, 2012).

Generally, DEIsoM ranks genes/isoforms highly associated with HCC on the top. In the top 10 list (the first four columns in Fig. 8), 60% of genes identified by DEIsoM as having DE isoforms are experimentally proven HCC biomarkers (Category 3), and 90% are HCC biomarkers plus DE genes verified *in vivo* (Category 3 + 2). On the contrary, BitSeqVB, RSEM-EBSeq, and Cuffdiff show a lower performance than DEIsoM 30 to 40% of genes having DE isoforms that are experimentally proved HCC biomarkers (Category 3), and 40% to 50% are HCC biomarkers plus DE genes verified *in vivo* (Category 3 + 2).

Even if we expand this search to top 50 lists (the fifth column in Fig. 8 and Supplementary Table S4), DEIsoM still identifies 18 genes (36%) as HCC biomarkers, and 10 genes (20%) as DE genes verified *in vivo*. However, BitSeqVB, RSEM-EBSeq, and Cuffdiff identify fewer literature proven DE genes than DEIsoM in the top 50 list (the last three columns in Fig. 8 and Supplementary Tables S5–7). BitSeqVB identifies 16 genes (32%) as HCC biomarkers, 12 genes (24%) as DE genes *in vivo*; RSEM-EBSeq identifies 12 genes (24%) as HCC biomarkers and 3 genes (6%) as DE genes verified *in vivo*; Cuffdiff identifies 11 genes (22%) as HCC biomarkers, 12 genes (24%) as DE genes *in vivo*. Therefore, DEIsoM has a clear superior ability to select DE genes that are supported by the published literature.

Moreover, the isoforms of four genes (FGFR2, survivin, ADAMTS13 and CD44) identified as DE by DEIsoM have been found to be up or down-regulated in HCC. This provides additional support for DE isoforms identified by DEIsoM. In the case of FGFR2 (ranked 62 of 11 950 genes), the FGFR2-IIIb isoform is down-regulated and has been related to HCC aggressive growth, while the FGFR2-IIIc isoform is expressed at the same level in normal and HCC tissues (Amann *et al.*, 2010). All three isoforms of survivin (ranked 120 of 11950 genes), survivin normal, survivin 2B and survivin Delta Ex3 have been detected in well, moderately and poorly differentiated HCC but none of these are found in normal tissues (Takashima *et al.*, 2005). RT-PCR results are available for ADAMTS13 (ranked 201 of 11950 genes) showing differences in the expression of three known isoforms (WT and 1, 2) between normal liver tissue and hepatoma cell lines (Shomron *et al.*, 2010). For CD44 (ranked 607 of 11950 genes), CD44-v6 is up-regulated in HCC, while CD44 standard form remains stable (Zhang *et al.*, 2010).

To more clearly understand the performance of different methods, we also examine the overlapping DE genes in the top 200 lists from the compared methods. Supplementary Table S2 shows the overlapping DE genes by feeding the FPKM of all isoforms from each method to EBSeq. This tests the quantification similarity between any two methods. According to the number of overlapping DE genes, the quantification performance of RSEM and BitSeqVB are the most similar, followed by RSEM and DEIsoM. Supplementary Table S3 shows the overlapping DE genes using the DE evaluation methods of their own. This tests the performance of both the quantification and DE identification. After changing the DE evaluation method, the number of overlapping DE genes between RSEM and BitSeqVB decreases from 96 to 62, while this number between RSEM and DEIsoM decreases from 74 to 14, which suggests that KL divergence performs differently from PPDE or PPLR. PPDE and PPLR are only sensitive to the absolute abundance change of an isoform, while KL divergence is sensitive to the overall isoform fractional pattern change within a gene, not limited to the absolute abundance change. This is useful in searching isoform switching events in many cases.

5 Discussion

In contrast to the models that treat each biological replicate separately, DEIsoM incorporates all biollogical replicates in one seamless framework. By capturing the shared information across multiple biological replicates, DEIsoM achieves a higher prediction accuracy and inter-replicate consistency than the alternative methods in the simulation studies (Section 3.1, 3.3, 3.4). This shared information comes from the intrinsic fact that all the replicates in one condition share the same underlying biological mechanism. As described in model construction (Section 2.1), we use a Dirichlet prior to represent a base fraction, which is characterized by the hyperparameter α and learned from data, and then sample the instance-specific fraction for each replicate. The fractions for different replicates are not necessarily the same, because we allow some within-condition variance, however, those fractions retain underlying coherence since they are sampled from the same Dirichlet prior (or the base fraction). In addition, as the conjugate prior for the multinomial distribution, the Dirichlet prior enables close form, efficient updates in our VB inference, which greatly benefits the computation. Furthermore, faster computing speed is gained using the VB algorithm, instead of the MCMC sampling used in MISO, during the inference step. The VB method converts a sampling problem to an optimization problem and speeds up the estimation (Section 3.2). DEIsoM is also promising in real applications. On the HCC dataset, by PCA plotting, we find that the normal and tumor samples can be linearly separated by the estimated expression levels of PCR verified DE genes, suggesting an accurate quantification of DE isoforms in DEIsoM. Using read coverage visualization, we find that the DEIsoM KL divergence is capable of identifying isoforms whose read coverage patterns change, and does not give false positive results for isoforms with low read abundance in both

conditions. This property is desirable in practice, since a low number of reads causes a large uncertainty in estimation. In DEIsoM, the posterior distributions of both conditions are close to the uniformly distributed prior if the read number is low, which reduces the KL divergence between the two conditions. However, neither Cuffdiff nor RSEM-EBSeq will automatically prune such isoforms (Section 4.2, 4.3). Moreover, a great number of isoforms predicted to be DE by DEIsoM are supported by the biological literature, providing encouraging results for real applications.

However, there are still some improvements that could be incorporated into DEIsoM. First, DEIsoM builds on the approach of MISO, which considers the quantification of isoforms gene by gene. In order to handle the reads multi-mapped to different gene loci, we have also added a variant version of DEIsoM that simultaneously considers all transcript isoforms, rather than performing a gene by gene analysis. This enhancement will allow the inclusion of multiply mapped reads into the analysis. However, the KL divergence is not applicable to this version, since KL divergence measures the isoform pattern change within a gene. Second, the KL divergence as a DE evaluation method is not based on a hypothesis test, but rather on the difference of the posterior distributions of fractional isoform expression between two conditions, so it only provides a rank instead of *P*-values to infer 'significantly' DE genes. However, KL divergence is sensitive to the overall isoform pattern change within a gene, and more differentiable for ranking isoforms/genes than *P*-values, which tend to give the same rank to many genes. DEIsoM allows the estimated isoform levels to be reported as FPKM, thereby allowing *P*-values to be calculated by many existing differential expression analysis methods. Lastly, DEIsoM assumes a known reference genome/transcriptome and the uniform read distribution. The misannotation or the non-uniformity of the read data may compromise the estimation accuracy in DEIsoM. We are considering including the novel isoform construction and the modeling of non-uniformly distributed read data into our future versions.

6 Conclusion

We propose a hierarchical Bayesian model, DEIsoM, for detecting DE isoforms using multiple biological replicates from two conditions. DEIsoM captures the information shared across replicates, and provides fast and accurate prediction compared to alternative methods in simulations. On the HCC real dataset, the estimated expression levels of PCR verified DE genes can be used as features to separate the tumor samples from their matched normal samples in PCA plots; read coverage visualization confirms that DEIsoM KL divergence is capable of identifying DE isoforms. DEIsoM is relatively resistant, compared to alternative methods, to identifying isoforms with low read abundance in both conditions as DE. Biological literature review suggests that the DE isoforms selected by DEIsoM have high relevance to HCC.

Acknowledgements

We would like to thank Eli Lilly and company for sharing the HCC data and providing the very helpful discussions.

Funding

This work was supported by NSF CAREER Award IIS-1054903, and the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Conflict of Interest: none declared.

References

Amann,T. *et al.* (2010) Reduced expression of fibroblast growth factor receptor 2IIIb in hepatocellular carcinoma induces a more aggressive growth. *Am. J. Pathol.*, **176**, 1433–1442.

Berasain,C. *et al.* (2010) Impairment of pre-mRNA splicing in liver disease: mechanisms and consequences. *World J. Gastroenterol.*, **16**, 3091–3102.

Bishop,C. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Blei,D.M. *et al.* (2003) Latent Dirichlet Allocation. *J. Mach. Learn. Res*, **3**, 993–1022.

Consortium,G. (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Dong,H. *et al.* (2009) Gene expression profile analysis of human hepatocellular carcinoma using sage and longsage. *BMC Med. Genomics*, **2**, 5.

Gierliński,M. *et al.* (2015) Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*.

Glaus,P. *et al.* (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**, 1721.

Grant,G.R. *et al.* (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.

Hensman,J. *et al.* (2015) Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, **31**, 3881.

Huang,Y. *et al.* (2017) Identification and functional analysis of differentially expressed genes in poorly differentiated hepatocellular carcinoma using RNA-seq. *Oncotarget*.

Jang,J.W. *et al.* (2012) Serum interleukin-6 and C-reactive protein as a prognostic indicator in hepatocellular carcinoma. *Cytokine*, **60**, 686–693.

Jordan,M.I. *et al.* (1999) An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

Kakaradov,B. *et al.* (2012) Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Bioinform.*, **13**(Suppl 6), S11.

Kan,Z. *et al.* (2013) Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.*, **23**, 1422–1433.

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

Langmead,B. and Salzberg,L.S. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Le,H.S. *et al.* (2013) Probabilistic error correction for RNA sequencing. *Nucleic. Acids. Res.*, **41**, e109.

Leng,N. *et al.* (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035.

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

Martin,J.A., and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Minka,T.P. (2000). Estimating a Dirichlet distribution. Technical report, M.I.T.

Nowicka,M. and Robinson,M. (2016) DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; referees: 2 approved]. *F1000 Res.*, **5**, 1356.

Ozsolak,F., and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Qiu,J.G. *et al.* (2008) Screening and detection of portal vein tumor thrombi-associated serum low molecular weight protein biomarkers in human hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.*, **134**, 299–305.

Ronning,G. (1989) Maximum likelihood estimation of Dirichlet distributions. *J. Statist. Comput. Simul.*, **34**, 215–221.

Shen,S. *et al.* (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl. Acad. Sci. USA*, **111**, E5593.

Shomron,N. *et al.* (2010) A splice variant of ADAMTS13 is expressed in human hepatic stellate cells and cancerous tissues. *Thromb. Haemost.*, **104**, 531–535.

Stegle,O. *et al.* (2010) Statistical tests for detecting differential RNA-transcript expression from read counts. *Nat. Proc.*

Sturgill,D. *et al*. (2013) Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-seq data with the splicing analysis kit (Spanki). *BMC Bioinform*., **14**, 320.

Sung,W.K. *et al*. (2012) Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet*., **44**, 765–769.

Takashima,H. *et al*. (2005) In vivo expression patterns of survivin and its splicing variants in chronic liver disease and hepatocellular carcinoma. *Liver Int*., **25**, 77–84.

Tan,G.S. *et al*. (2014) Novel proteomic biomarker panel for prediction of aggressive metastatic hepatocellular carcinoma relapse in surgically resectable patients. *J. Proteome Res*., **13**, 4833–4846. PMID: 24946162.

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

Trapnell,C. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., **7**, 562–578.

Vaquero-Garcia,J. *et al*. (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, e11752.

Wang,F. *et al*. (2015) Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). *Med. Oncol*., **32**.

Wang,F. *et al*. (2017) A transcriptome profile in hepatocellular carcinomas based on integrated analysis of microarray studies. *Diagn. Pathol*., **12**, 4.

Wang,K. *et al*. (2009) Identification of tumor-associated antigens by using SEREX in hepatocellular carcinoma. *Cancer Lett*., **281**, 144–150.

Yim,S.H. and Chung,Y.J. (2010) An overview of biomarkers and molecular signatures in HCC. *Cancers*, **2**, 809–823.

Zhang,T. *et al*. (2010) PCBP-1 regulates alternative splicing of the CD44 gene and inhibits invasion in human hepatoma cell line HepG2 cells. *Mol. Cancer*, **9**.