This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

A generalized dSpliceType framework to detect differential splicing and differential expression events using RNA-Seq

Dongxiao Zhu[1], Nan Deng[1,2], Changxin Bai[1]

[1]Department of Computer Science, Wayne State University, Detroit, MI 48202 USA
[2]Biostatistics and Bioinformatics Research Center, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048 USA

Transcriptomes are routinely compared in term of a list of differentially expressed genes followed by functional enrichment analysis. Due to the technology limitations of microarray, the molecular mechanisms of differential expression is poorly understood. Using RNA-seq data, we propose a generalized dSpliceType framework to systematically investigate the synergistic and antagonistic effects of differential splicing and differential expression. We applied the method to two public RNA-seq data sets and compared the transcriptomes between treatment and control conditions. The generalized dSpliceType detects and prioritizes a list of genes that are differentially expressed and/or spliced. In particular, the multivariate dSpliceType is among the fist to utilize sequential dependency of normalized base-wise read coverage signals and capture biological variability among replicates using a multivariate statistical model. We compared dSpliceType with two other methods in terms of five most common types of differential splicing events between two conditions using RNA-Seq. dSpliceType is free available from http://orleans.cs.wayne.edu/dSpliceType/

Index Terms—**Differential Splicing, Differential Expression, Multivariate Statistical Models, RNA-Seq, Transcriptomics.**

I. Introduction

A Messenger RNA (mRNA) transcriptome, henceforth referred to as transcriptome, is composed of a set of all mRNA molecules produced in one cell or a population of cells. It is highly dynamic (changes over time points and conditions) and complex (mainly due to alternative splicing). In comparison to genomes and proteomes, transcriptomes offer a unique angle to uncover molecular and pathway mechanisms of the phenotypic traits. The convenience and power of studying transcriptomes is particularly pronounced for plants and amphibians with large and complex genomes composed of an excessive amount of repetitive DNA elements. Transcriptomes are critical molecular drivers of phenotypes in plant, animal and microbe.

Hence comparing and deciphering transcriptomes in a variety of organisms are a high priority in modern biomedical research. Our capability of comparing and deciphering transcriptomes quickly evolves with the development of high throughput transcriptome profiling techniques. Twenty years ago, transcriptome analysis is largely based on microarray data. Using microarray technology, gene expression signal was analogously captured by hybridizing the biotin-labeled mRNA transcript fragments to a set of DNA probes printed on the gene chip and measuring the fluorescent intensities of each individual probe or probe set. In response to the computational challenges brought by high throughput gene expression data, an array of excellent computational tools were developed to prioritize the differentially expressed genes. Earlier approaches are based on statistical testing whether the mean expression values in each condition are equal or not [1]. Due to the small sample size of microarray data, variance among the replicated samples are often not stably estimated. Latter approaches strived to improve the variance estimate, some of more effective methods include Significant Analysis of Microarray [2] and Empirical Bayes [3]. Consequently, the transcripomes were compared based on a list of top ranked differentially expressed genes, and were deciphered by examining the functional and/or pathway enrichment of the gene list. Limited by the resolution of microarray technology, i.e, only dozens of probes were designed to interrogate the expression of each gene, the underlying molecular mechanisms driving the observed differential expression remain opaque, for example, on the specific exon that is differentially expressed.

Next-generation sequencing technology makes it possible for an in-depth dissection of molecular mechanism of differential expression. After aligning the reads to the reference genome, the pile-up data essentially yields expression signal at base-wise resolution, i.e., it gives information on how many copies of each base are transcribed in the transcriptome. These base-wise expression signals, on one hand, permit a more powerful detection of differential expression due to a massive amount of data, on the other hand, make it possible to zoom in and figure out the specific regions of the genes that are differentially expressed, the latter is often represented by differential alternative splicing or differential splicing.

Alternative splicing plays a key role in contributing to the transcriptome diversity in eukaryotes [4]. It occurs in more than 90% of human genes in different types [5], including skipped exon (SE), retained intron (RI), alternative 3' or 5' splice sites (A3SS or A5SS), and mutually exclusive exons (MXE) [4]. Studies have shown that differential splicing, for example, dysregulation of alternative splicing events, may lead to malignant phenotypes, such as human disease [4], [5], [6], [7].

Driven by the technology development, numerous methods have been developed to detect and prioritize differential splic-

Corresponding Author: D. Zhu (email: dzhu@wayne.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

ing. One type of existing methods are based on the estimation of full-length transcripts, and to estimate relative transcript abundances followed by a statistical test of relative abundances within a gene between conditions to quantify the differences. This type of methods, such as Cufflinks/Cuffdiff [8] and others [9], [10], [11], is powerful. However, they rely on accurate estimation of transcript relative abundances, which is a non-trivial problem. Another type of methods detect differentially spliced genes by comparing read counts either on all exons within a gene, such as SplicingCompass [12] and FDM [13], or on a single exon, e.g. DEXSeq [14]. These methods can potentially detect differentially spliced genes, but can not specify the regions or associated types of differential splicing events. Newer methods are event-based, directly detecting differential splicing events. Certain methods, such as MISO [15] and SpliceTrap [16], focus on the detection of SE events. More recently, MATS [17] and DiffSplice [18] are capable of detecting multiple types of events. However, either MCMC method (MATS) or permutation test (DiffSplice) makes the detection of differential splicing events time consuming. Our previous work, univariate dSpliceType [19] and multivariate dSpliceType [20], efficiently detects five types of different splicing events using closed-form solutions.

Albeit the existing comparison methods to detect differential expression and differential splicing are useful, they may fail to detect synergistic and antagonistic effects exist between differential splicing and differential expression events. For example, differential splicing can either strengthen or weaken differential expression. Studying these effects will lead to a deep understanding of molecular mechanisms of the phenotype changes. Here we present an integrated methodology framework to detect various types of differential splicing and differential expression events using RNA-Seq. The generalized dSpliceType is among the first to detect both differential expression and differential splicing events using base-wise expression signals. It utilizes sequential dependency of normalized base-wise read coverage signals and captures biological variability among replicated samples using multivariate statistical models. Using both simulation and real-world data, we demonstrate the advantages of the generalized dSpliceType and compared to the selected methods. The rest of the paper is organized as follows. Section 2 presents the descriptions of concepts and methods for prioritizing differential expression and differential splicing. Section 3 presents the experimental results and comparison, and Section 4 concludes the paper, summarizes the main points and discusses the future outlook.

II. Methods

The generalized dSpliceType is a parametric statistical framework for detecting both differential expression and differential splicing events using RNA-Seq data. Figure 1 shows the workflow of dSpliceType, and the detailed procedure is described as follows.

A. Detecting and prioritizing differential splicing events

1) Extracting candidate splicing events

dSpliceType extracts candidate splicing events for the five most common types of alternative splicing from gene annotation database along with supported junction reads as shown in Figure 1A and Figure 1B. With intron removal, candidate splicing events consist of concatenating left common exon, spliced exon(s) (for SE and MXE events) or exonic region (for RI, A3SS and A5SS events) and right common exon. Two spliced exons are for MXE event. The detailed strategies for extracting different types of candidate splicing events are described in [19]. Novel candidate splicing events can be extracted by incorporating novel junction reads.

2) Calculating normalized logRatio of RNA-Seq splicing indexes

After extracting candidate splicing events, the read coverage signal (Figure 1C) and the RNA-Seq splicing index (Figure 1D.1) at each nucleotide location are calculated in terms of differential splicing for each replicate in both conditions. The RNA-Seq splicing index at the *i*th nucleotide location is denoted as SI_i. The read coverage signal c_i is normalized by read coverage signals on the two common exons of the candidate splicing event (exons in black color as shown in Figure 1B) as

$$SI_i = \frac{c_i}{\frac{\sum_{p=1}^{le_l} c_p + \sum_{q=1}^{le_r} c_q}{le_l + le_r}},$$

in which le_l and le_r are the length of the left and the right common exons, respectively. The formula of calculating SI_i using local normalizer is initially given in [19], and a similar splicing index has been used for analysis of differential splicing in microarray studies [21].

After that, the logRatio of normalized RNA-Seq splicing indexes (Figure 1E) on each nucleotide of a candidate splicing event between two conditions is calculated. We denote it as $\log\left(SI_{\text{caseSample}_{im}} / \overline{SI_{\text{controlSample}_i}}\right)$, where m is the index of replicates in case condition and $\overline{SI_{\text{controlSample}_i}}$ is the average of RNA-Seq splicing indexes at ith nucleotide location of replicates in control condition.

Since the sequencing and alignment biases are more likely to affect read coverage signals at the same nucleotide locations on all samples in the same way, the effect of biases from RNA-Seq is substantially reduced by taking ratio of normalized RNA-Seq splicing indexes at each nucleotide location of replicates in two conditions.

3) The multivariate conditional normal distribution model for the normalized logRatio of RNA-Seq splicing indexes

As shown in Figure 1E, we denote the normalized $\log\left(SI_{\text{caseSample}_{im}} / \overline{SI_{\text{controlSample}_i}}\right)$ at the ith nucleotide along the candidate splicing event as $\mathbf{X}_i$, which is a m-dimensional normal random vector from $N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, for *i* = 1,...,*n*. For computational simplicity, we capture the sequential dependency between $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$, which follows [22]:

$$\mathbf{X}_i | \mathbf{X}_{i-1} = \mathbf{x}_{i-1} \sim N_m(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}),$$

where

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu_i} + \boldsymbol{\Sigma}_{i,i-1} \boldsymbol{\Sigma}_{i-1,i-1}^{-1} (\mathbf{x}_{i-1} - \boldsymbol{\mu}_{i-1}),$$

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}_{i,i-1} \boldsymbol{\Sigma}_{i-1,i-1}^{-1} \boldsymbol{\Sigma}_{i-1,i}.$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

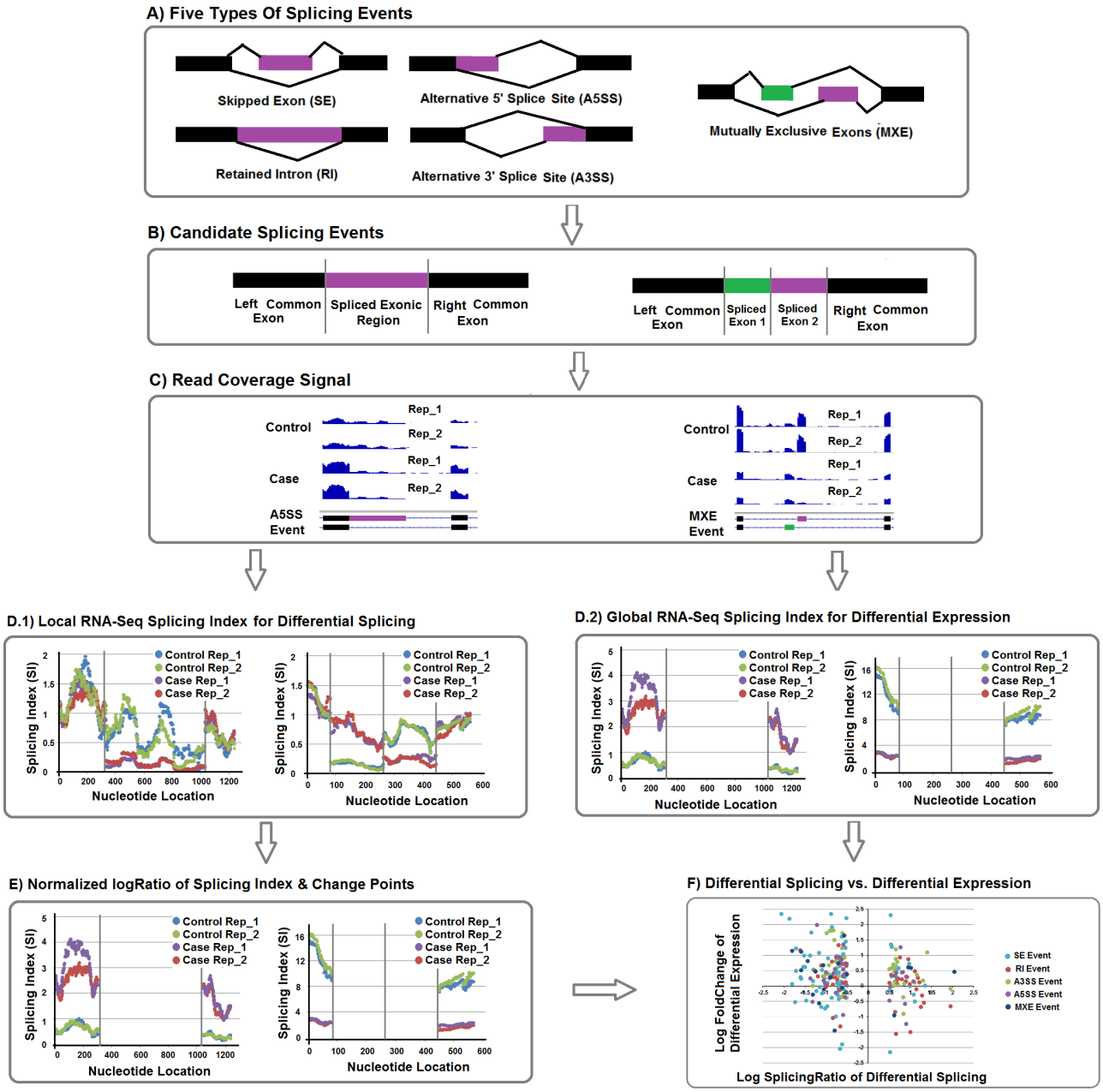

Figure 1. The workflow of the generalized dSpliceType for detecting various types of differential splicing events and differential expression. A) Five most common types of splicing events. Left panel represents SE, RI, A3SS and A5SS events, and right panel represents MXE event. B) Candidate splicing events are compiled by removing introns and concatenating left common exon, spliced exon(s) or exonic region and right common exon. C) For each candidate splicing event (illustrated by A5SS and MXE events), read coverage signals are calculated on nucleotides for each replicate in both conditions. D.1) and E) Local RNA-Seq splicing indexes and normalized logRatio of splicing indexes are calculated based on read coverage signals. dSpliceType detects the differential splicing events by identifying change points on the ending locations of exon(s) or exonic region D.2) Global RNA-Seq splicing indexes are calculated on left and right common exons. dSpliceType detects differential expression using a multivariate or a univariate statistical test based on the data quality. F) A scatter-plot of differential expression against differential splicing to detect their synergistic and antagonistic effects.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

$\mathbf{\Sigma}_{i,i}$ and $\mathbf{\Sigma}_{i-1,i-1}$ represent the variances of $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$, respectively, while $\mathbf{\Sigma}_{i,i-1}$ and $\mathbf{\Sigma}_{i-1,i}$ represent the covariances between $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$ and it's transpose. $\mathbf{\Sigma}^{-1}_{i-1,i-1}$ represents the generalized inverse of $\mathbf{\Sigma}_{i-1,i-1}$. The sequence of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ can be considered as a series of multivariate conditional normal random variables from $N_m(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{\Sigma}})$, for $i = 2, ..., n$, where n is the total exonic length of the candidate splicing event. If no differential splicing happens, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{\Sigma}}$ are assumed to be constant mean vector of $\boldsymbol{\mu}$ and covariance matrix of $\mathbf{\Sigma}$; while deviations from the constant mean vector and covariance matrix in the spliced region may indicate a differential splicing event.

4) The hypothesis testing

The identification of differential splicing event among multiple samples can be transformed to identify multiple change points at exon boundaries according to different types of candidate splicing events, and can be further defined as testing the null hypothesis for both mean and covariance parameters in the series of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ [23]:

$$\begin{aligned} \mathrm{H}_0 : \tilde{\boldsymbol{\mu}}_1 = \tilde{\boldsymbol{\mu}}_2 = ... = \tilde{\boldsymbol{\mu}}_n = \boldsymbol{\mu} \text{ and} \\ \tilde{\mathbf{\Sigma}}_1 = \tilde{\mathbf{\Sigma}}_2 = ... = \tilde{\mathbf{\Sigma}}_n = \mathbf{\Sigma}. \end{aligned} \quad (1)$$

For SE, A3SS, A5SS and RI, the alternative hypothesis is:

$$\begin{aligned} &\mathrm{H}_1 : \tilde{\boldsymbol{\mu}}_1 = ... = \tilde{\boldsymbol{\mu}}_i \neq \tilde{\boldsymbol{\mu}}_{i+1} = ... = \tilde{\boldsymbol{\mu}}_j \neq \tilde{\boldsymbol{\mu}}_{j+1} = ... = \tilde{\boldsymbol{\mu}}_n \text{ and} \\ &\tilde{\mathbf{\Sigma}}_1 = ... = \tilde{\mathbf{\Sigma}}_i \neq \tilde{\mathbf{\Sigma}}_{i+1} = ... = \tilde{\mathbf{\Sigma}}_j \neq \tilde{\mathbf{\Sigma}}_{j+1} = ... = \tilde{\mathbf{\Sigma}}_n, \end{aligned} \quad (2)$$

where i and j, $1 < i < j < n$, are the ending locations of the left common exon (in black) and the spliced exon/exonic region (in purple), respectively, as shown on the left panel of Figure 1B. For each candidate splicing event of the four types, a significant differential splicing event is detected when the null hypothesis (1) is rejected at a given significance level α.

For MXE, the alternative hypothesis is:

$$\begin{aligned} &\mathrm{H}_1 : \tilde{\boldsymbol{\mu}}_1 ... = \tilde{\boldsymbol{\mu}}_i \neq \tilde{\boldsymbol{\mu}}_{i+1} ... = \tilde{\boldsymbol{\mu}}_j \neq \tilde{\boldsymbol{\mu}}_{j+1} ... = \tilde{\boldsymbol{\mu}}_k \neq \tilde{\boldsymbol{\mu}}_{k+1} ... = \tilde{\boldsymbol{\mu}}_n \\ &\text{and} \\ &\tilde{\mathbf{\Sigma}}_1 ... = \tilde{\mathbf{\Sigma}}_i \neq \tilde{\mathbf{\Sigma}}_{i+1} ... = \tilde{\mathbf{\Sigma}}_j \neq \tilde{\mathbf{\Sigma}}_{j+1} ... = \tilde{\mathbf{\Sigma}}_k \neq \tilde{\mathbf{\Sigma}}_{k+1} ... = \tilde{\mathbf{\Sigma}}_n, \end{aligned} \quad (3)$$

where i, j and k, $1 < i < j < k < n$, are the ending locations of the left common exon (in black) and the two spliced exons (in purple and green), respectively, as shown on the right panel of Figure 1B. A significant differential splicing MXE event is detected when the null hypothesis (1) is rejected at a given significance level α.

5) The Schwarz information criterion

In order to test the null hypothesis (1) against the alternative hypothesis (2) or (3), the Schwarz information criterion (SIC)-based method [24] is employed. The smaller SIC score indicates the better data fitting of a model. Thus, the hypothesis testing can be converted into selecting a model such that the null hypothesis (1) represents a model without change of mean and covariance parameters, while the alternative hypothesis (2) or (3) represents models with different means and covariances specified by two or three change points. Since, on average, more than 100 nucleotides are in the common and spliced exons/exonic regions, number of $\mathbf{X}_i$'s are considered to be sufficient for estimating model parameters and calculating SIC scores.

We denote SIC(n) as the SIC corresponding to the null hypothesis (1), which is derived as :

$$SIC(n) = -2 \log L_0(\hat{\tilde{\boldsymbol{\mu}}}, \hat{\tilde{\mathbf{\Sigma}}}) + \frac{m(m+3)}{2} \log n,$$

where the log likelihood is

$$\log L_0(\hat{\tilde{\boldsymbol{\mu}}}, \hat{\tilde{\mathbf{\Sigma}}}) = -\frac{1}{2} mn \log 2\pi - \frac{n}{2} \log |\hat{\tilde{\mathbf{\Sigma}}}| - \frac{n}{2}.$$

So, we have

$$SIC(n) = mn \log 2\pi + n \log |\hat{\tilde{\mathbf{\Sigma}}}| + n + \frac{m(m+3)}{2} \log n.$$

$\log L_0(\hat{\tilde{\boldsymbol{\mu}}}, \hat{\tilde{\mathbf{\Sigma}}})$ is the maximum log likelihood function, and $\hat{\tilde{\boldsymbol{\mu}}}$ and $\hat{\tilde{\mathbf{\Sigma}}}$ are MLEs of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{\Sigma}}$ under H_0, respectively, in which

$$\hat{\tilde{\mathbf{\Sigma}}} = \hat{\mathbf{\Sigma}}_{i,i} - \hat{\mathbf{\Sigma}}_{i,i-1} \hat{\mathbf{\Sigma}}^{-1}_{i-1,i-1} \hat{\mathbf{\Sigma}}_{i-1,i},$$

where

$$\hat{\mathbf{\Sigma}}_{i,i} = \frac{1}{n-1} \sum_{i=2}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_i)(\mathbf{X}_i - \overline{\mathbf{X}}_i)',$$

$$\hat{\mathbf{\Sigma}}_{i-1,i-1} = \frac{1}{n-1} \sum_{i=2}^{n} (\mathbf{X}_{i-1} - \overline{\mathbf{X}}_{i-1})(\mathbf{X}_{i-1} - \overline{\mathbf{X}}_{i-1})',$$

$$\hat{\mathbf{\Sigma}}_{i,i-1} = \frac{1}{n-1} \sum_{i=2}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_i)(\mathbf{X}_{i-1} - \overline{\mathbf{X}}_{i-1})',$$

$$\hat{\mathbf{\Sigma}}_{i-1,i} = \hat{\mathbf{\Sigma}}'_{i,i-1},$$

$$\overline{\mathbf{X}}_i = \frac{1}{n-1} \sum_{i=2}^{n} \mathbf{X}_i,$$

$$\overline{\mathbf{X}}_{i-1} = \frac{1}{n-1} \sum_{i=2}^{n} \mathbf{X}_{i-1}.$$

Corresponding to $\mathrm{H}_1(2)$ with two change points *i* and *j*, the SIC for differential splicing events (SE, RI, A3SS and A5SS), denoted by SIC(*i*, *j*) for fixed i and $j, m \leqslant i, j \leqslant n - m$, is derived as:

$$\begin{aligned} SIC(i,j) &= -2 \log L_1(\hat{\tilde{\boldsymbol{\mu}}}_1, \hat{\tilde{\boldsymbol{\mu}}}_2, \hat{\tilde{\boldsymbol{\mu}}}_3, \hat{\tilde{\mathbf{\Sigma}}}_1, \hat{\tilde{\mathbf{\Sigma}}}_2, \hat{\tilde{\mathbf{\Sigma}}}_3) \\ &\quad + \frac{3m(m+3)}{2} \log n \\ &= mn \log 2\pi + i \log |\hat{\tilde{\mathbf{\Sigma}}}_1| + (j-i) \log |\hat{\tilde{\mathbf{\Sigma}}}_2| \\ &\quad + (n-j) \log |\hat{\tilde{\mathbf{\Sigma}}}_3| + n + \frac{3m(m+3)}{2} \log n. \end{aligned}$$

Estimating $\hat{\tilde{\mathbf{\Sigma}}}_1$, $\hat{\tilde{\mathbf{\Sigma}}}_2$ and $\hat{\tilde{\mathbf{\Sigma}}}_3$ are similar to estimating $\hat{\tilde{\mathbf{\Sigma}}}$, except for using intervals $[1, i]$, $[i+1, j]$ and $[j+1, n]$ instead of using interval $[1, n]$ accordingly.

Similarly, corresponding to $\mathrm{H}_1(3)$ with three change points *i*, *j* and *k*, the SIC for differential splicing events of MXE ,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

denoted by SIC(i, j, k) for fixed i, j and $k, m \leqslant i,j,k \leqslant n-m$, is derived as:

$$\begin{aligned} SIC(i,j,k) &= -2\log L_2(\hat{\tilde{\boldsymbol{\mu}}}_1, \hat{\tilde{\boldsymbol{\mu}}}_2, \hat{\tilde{\boldsymbol{\mu}}}_3, \hat{\tilde{\boldsymbol{\mu}}}_4, \hat{\tilde{\boldsymbol{\Sigma}}}_1, \hat{\tilde{\boldsymbol{\Sigma}}}_2, \hat{\tilde{\boldsymbol{\Sigma}}}_3, \hat{\tilde{\boldsymbol{\Sigma}}}_4) \\ &\quad + 2m(m+3)\log n \\ &= mn\log 2\pi + i\log|\hat{\tilde{\boldsymbol{\Sigma}}}_1| + (j-i)\log|\hat{\tilde{\boldsymbol{\Sigma}}}_2| \\ &\quad + (k-j)\log|\hat{\tilde{\boldsymbol{\Sigma}}}_3| + (n-k)\log|\hat{\tilde{\boldsymbol{\Sigma}}}_4| \\ &\quad + n + 2m(m+3)\log n. \end{aligned}$$

Estimating $\hat{\tilde{\boldsymbol{\Sigma}}}_1$, $\hat{\tilde{\boldsymbol{\Sigma}}}_2$, $\hat{\tilde{\boldsymbol{\Sigma}}}_3$ and $\hat{\tilde{\boldsymbol{\Sigma}}}_4$ are similar to estimating $\hat{\tilde{\boldsymbol{\Sigma}}}$, except for using intervals $[1,i]$, $[i+1,j]$, $[j+1,k]$ and $[k+1,n]$ instead of interval $[1,n]$ accordingly.

According to the principle of information criterion [23], the null model fits the data in the sequence of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ better if

$$SIC(n) < SIC(i,j) \text{ or } SIC(n) < SIC(i,j,k).$$

Otherwise, the model with two change points better fits the data in the sequence of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ for differential splicing events SE, A3SS, A5SS and RI, and the change points i and j are at the ending locations of the left common exon and the spliced exon or exonic region.

Similarly, the model with three change points better fits the data in the sequence of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ for differential splicing event MXE, and the change points i, j and k are the ending locations of the left common exon and the two spliced exons.

6) The test statistic

According to [23], the difference between the SIC scores of the models with and without change points,

$$\Delta_n = SIC(i,j) - SIC(n) \text{ and } \Delta_n = SIC(i,j,k) - SIC(n),$$

can be used as a statistic, and we use the asymptotic null distribution of Δ_n to calculate the approximate p-value for the test of the null hypothesis (1) against the alternative hypothesis (2) or (3) as

$$p\text{-value} = 1 - exp\left\{-2exp[b_{2m}(\log n) - a(\log n)\lambda_n^{1/2}]\right\},$$

where

$$\lambda_n = 2\log n - \Delta_n,$$

$$a(\log n) = (2\log\log n)^{1/2},$$

$$b_{2m}(\log n) = 2\log\log n + m\log\log\log n - \log\Gamma(m),$$

$$\Gamma(m) = (m-1)!.$$

The raw p-values of the multiple tests are adjusted using the stringent Bonferroni's procedure.

B. Detecting and prioritizing differential expression

Using the base-wise expression signals derived from Section II-A, we denote $SI_{\text{controlSample}_{im}}$ and $SI_{\text{caseSample}_{im}}$ using global normalizer on the left and right common exons at the ith nucleotide along the candidate splicing event as the values of $\mathbf{X}_0$ and $\mathbf{X}_1$, which are m-dimensional normal random vectors from $N_m(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. μ_{0j} and μ_{1j}, where $j = 1, ..., m$, represent the means of jth series of base-wise expression signals in control and case conditions respectively. We consider the following hypothesis test:

$$\mathrm{H}_0 : \mu_{01} = \mu_{11} \text{ and } \mu_{02} = \mu_{12} \text{ and } \ldots \text{ and } \mu_{0m} = \mu_{1m} \quad (1)$$

against

$$\mathrm{H}_1 : \mu_{0j} \neq \mu_{1j} \text{ for at least one } j, j = 1, \ldots, m. \quad (2)$$

A significant differential expression is detected when the null hypothesis (1) is rejected at a given significance level α. To test the hypothesis, we calculate the Hotelling τ^2 as the test statistic below:

$$\tau^2 = n\mathbf{X}^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}\mathbf{X},$$

where

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{X_i}\mathbf{X_i^T}.$$

$\mathbf{S}$ is variance-covariance matrix that can be calculated from data. If we assume the different covariance matrices between case and control conditions, $\mathbf{S}$ will be calculated by a weighted form as below:

$$S = \frac{S_0}{n_0 - 1} + \frac{S_1}{n_1 - 1},$$

where S_0 and S_1 is the quantities $\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i^T}$ calculated in control and case conditions respectively, n_0 and n_1 are the corresponding sample (replicate) sizes.

For large n, τ^2 is approximately chi-square distributed with m degrees of freedom. p-value can therefore be calculated to access the statistical significance of differential expression.

Based on our previous studies on estimating correlation matrix from replicated data [25], it is worth mentioning that the above-mentioned multivariate statistical test works well with the 'noisy' data where moderate or poor correlations exist between the replicates. For 'clean' data with high correlation among the replicates, a more suitable approach would be averaging over the replicates in each condition and followed by a univariate statistical test such as t-test [1], SAM [2] or Empirical Bayes [3].

III. Results

A. Simulation studies

1) Simulation data sets

We evaluated the accuracy of dSpliceType [20] and compared the performance with two existing methods, MATS [17] and Cufflinks/Cuffdiff [8], using simulation studies. FluxSimulator [26] was used to simulate various splicing ratios of splicing events and generated 4 groups of RNA-Seq data sets on the entire human transcriptome. Each group includes 3 replicates in control and case conditions, respectively; and each replicate consists of 30 million, 50 million, 100 million and 200 million paired-end reads with 100bp in length for each group.

We mapped the simulated RNA-Seq data sets uniquely to the human reference genome (hg19/GRCh37) using Tophat2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

[27] and Bowtie2 [28]. To evaluate and compare the three methods, the alignment results in BAM format were served as inputs for the latest version of MATS (3.0.8) and Cufflinks/Cuffdiff (2.1.1) using default parameters. Read coverage signals (.bedgraph files) converted from alignment results (.bam files) using BEDtools [29] and read junctions (.bed files) were used as inputs for dSpliceType. The complete Ensembl annotation database and the significance level of 0.05 for adjusted *p*-values were used to detect differentially spliced genes for Cufflinks/Cuffdiff and differential splicing events for dSpliceType and MATS. To control false positives and biological significance of events, we further set parameters of dSpliceType such that the average read coverage on the spliced exonic region is more than 5, the average ratio of normalized RNA-Seq splicing indexes on the spliced exonic region is greater than 1.2 or smaller than 0.8. Please note that the detected differentially spliced genes in Table I and Figure 2 are all true positives, and no false positive is detected by all three methods with their parameter settings.

2) Simulation results of detecting differentially spliced genes

We compared the overall performances of the three computational methods on detecting differentially spliced genes in 4 groups of simulation data sets. We collected the differentially spliced genes directly from the result file of Cuffdiff (splicing.diff). For dSpliceType and MATS, a gene is considered to be differentially spliced if any type of differential splicing event was detected by the method for that gene. Table I shows that dSpliceType outperforms the other two methods by achieving the highest numbers and detection rates in all simulation data sets. One possible reason for the lowest detection rate of Cuffdiff is the challenges in estimating transcript relative abundances of genes with many annotated transcripts.

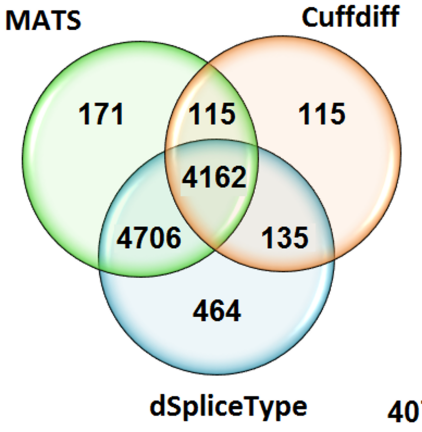

Figure 2. The comparison of the differentially spliced genes detected by dSpliceType, MATS and Cuffdiff (200M simulation data set).

To better evaluate the performance of dSpliceType, we compared the differentially spliced genes detected by the three methods in the 200 million simulated data set. As shown in Figure 2, there are 4,162 true differentially spliced genes detected by all the methods, and 464, 171 and 115 genes were exclusively detected by dSpliceType, MATS and Cuffdiff, respectively. We further examine the 464 genes detected exclusively by dSpliceType, and found that our method is able to better detect differentially spliced genes of low abundances ($0<$FPKM<1 and $1<$FPKM<5) than the competing methods as shown in Table II.

In addition to differentially spliced genes of relatively low abundances, a large number of differentially spliced genes detected by dSpliceType is overlapped with that detected by the other two methods as shown in Figure 2. Therefore, dSpliceType is demonstrated to be able to detect differentially spliced genes in a large dynamic range of expressed genes.

3) Simulation result of detecting differential splicing events

Since both dSpliceType and MATS are event-based methods, we focused on each type of splicing events to further investigate the differences between them. Table III shows that for each data set, dSpliceType outperforms MATS by 4% to 19% of detected rate on SE and MXE splicing events. For A3SS and A5SS events, MATS outperforms dSpliceType in some of the data sets while the rates of detection are still comparable. This is because when the spliced regions of A3SS or A5SS events are short, e.g., less than 5 nucleotides, data points used to estimate model parameters and calculate SIC scores may not sufficient. For RI event, MATS slightly outperforms dSpliceType in each data set. The possible reason is that since the spliced regions of RI event are usually longer than 1,000 nucleotides, more reads need to be sequenced to cover the long spliced regions, which can make model calculation more accurate. Therefore, when the number of reads reaches to 200 million, the detected rates of the two methods are quite close (76% of dSpliceType vs. 79% of MATS).

4) Runtime comparison

Table IV shows the runtime comparison of the three methods among the 4 groups of simulation data sets on the same Linux Ubuntu Server with 4 x Twelve-Core AMD Opteron 2.6GHz and 256GB RAM. For each data set, the runtime of dSpliceType is faster than the other two. The runtime of dSpliceType on each data set can be separated into two parts, the time of converting alignment results to read coverage signals using BEDtools [29] and the time of detecting differential splicing events by dSpliceType. The increase in number of reads reflects more of the increase in conversion time, not detection time.

B. Real-world data analysis

1) Human embryonic stem cell lines data set

a) RNA-Seq data and pre-processing: We applied dSpliceType to a public Illumina HiSeq 2000 paired-end RNA-Seq data set of human H1 and H1 derived neuronal progenitor cell lines (shorted as H1 and H1-npc). The data set can be accessed from NIH Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/) with NCBI SRA number SRR488684, SRR488685, SRR486241 and SRR486242 as two replicates of H1 and H1-npc cell lines, respectively. For each replicate, about 200 million reads (100bp $\times$ 2) were sequenced. For real-world RNA-Seq data analysis, the alignment procedure, the input files and parameters for dSpliceType are similar to simulation studies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

Table I
COMPARISON OF THE DIFFERENTIALLY SPLICED GENES DETECTED BY DSPLICETYPE, MATS AND CUFFDIFF IN 4 GROUPS OF SIMULATION DATA SETS. FOR EACH METHOD, THE HIGHEST DETECTION RATE IS IN BOLD FACE.

# of Reads	# of Spliced Genes[1]	Methods dSpliceType		MATS		Cuffdiff	
30M	10,275	8,054	78%	7,148	70%	2,086	20%
50M		8,701	85%	7,977	78%	2,626	26%
100M		9,170	89%	8,704	85%	3,425	33%
200M		9,467	**92%**	9,154	**89%**	4,527	**44%**

[1]The total number of differentially spliced genes in the simulation data sets.

Table II
THE PERCENTAGE OF DIFFERENTIALLY SPLICED GENES OF RELATIVELY LOW ABUNDANCES IN BOTH CONDITIONS DETECTED BY EACH METHOD EXCLUSIVELY AND ALL METHODS (200 MILLION SIMULATED DATA SET).

	# of Spliced Genes[1]	0 <FPKM <1	1 <FPKM <5
dSpliceType	464	11%	25%
MATS	171	5%	12%
Cuffdiff	115	0%	3%
All Methods	4,162	0%	1%

[1]The total number of differentially spliced genes detected by each method and all three methods.

Table III
COMPARISON OF THE DIFFERENTIAL SPLICING EVENTS DETECTED BY DSPLICETYPE AND MATS IN 4 GROUPS OF SIMULATED DATA SETS. FOR EACH METHOD IN EACH TYPE OF SPLICING EVENT, THE HIGHEST DETECTED RATE IS HIGHLIGHTED IN BOLD.

Type of Splicing	# of Splicing Events[1]	# of Reads	Methods dSpliceType		MATS	
SE	8,031	30M	5,853	**73%**	4,795	60%
		50M	6,341	**79%**	5,493	68%
		100M	6,706	**84%**	6,196	77%
		200M	6,880	**86%**	6,612	82%
A3SS	3,711	30M	2,567	**69%**	2,173	59%
		50M	2,758	**74%**	2,482	67%
		100M	2,914	**79%**	2,845	77%
		200M	3,007	81%	3,048	**82%**
A5SS	3,175	30M	2,150	**68%**	1,888	59%
		50M	2,356	**74%**	2,224	70%
		100M	2,489	78%	2,499	**79%**
		200M	2,559	81%	2,672	**84%**
RI	1,661	30M	728	44%	1,009	**61%**
		50M	901	54%	1,101	**66%**
		100M	1,092	66%	1,235	**74%**
		200M	1,260	76%	1,317	**79%**
MXE	1,366	30M	1,126	**82%**	855	63%
		50M	1,189	**87%**	956	70%
		100M	1,242	**91%**	1,051	77%
		200M	1,263	**92%**	1,107	81%

[1]For each type of splicing events, the number of differential splicing events in the simulated data sets. M stands for million.

Table IV
RUNTIME COMPARISON OF DSPLICETYPE, MATS AND CUFFDIFF IN 4 GROUPS OF SIMULATION DATA SETS. THE SHORTEST RUNTIMES ARE HIGHLIGHTED IN BOLD.

Methods	30M	50M	100M	200M
		(Hours : Minutes)		
BEDTools + dSpliceType	0:36+0:25	0:48+0:29	1:30+0:31	2:30+0:32
Total	**1:01**	**1:17**	**2:01**	**3:02**
Cuffdiff	2:52	3:01	3:31	4:38
MATS	17:02	19:13	30:35	40:41

dSpliceType (1 thread), Cuffdiff (6 threads), and MATS (1 thread). The runtime of Cuffdiff includes gene and transcript relative abundance estimation, differential expression analysis and differential splicing analysis. The runtime of MATS includes the conversion time from .bam to .sam, and differential splicing analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

b) Detection of differential splicing events with gene-level differential expression: dSpliceType detected amount of differential splicing events between H1 and H1-npc cell lines. We consider differential splicing events with biologically significance if $|\log \text{SplicingRatio}| > 0.5$. In Figure 3, we plot the detected differential splicing events vs. their gene-level differential expression using splicing ratio and fold change in log space.

For illustration purpose, in Figure 4 - 8, we present five differential splicing events detected by dSpliceType with different types of alternative splicing and gene-level differential expression. In each case study, the upper panel shows the read coverage signals of the differential splicing event; the lower panel includes the plots of local RNA-Seq splicing index for differential splicing, logRatio of splicing index and detected change points and global splicing index for differential expression, respectively. For these case studies, MATS can detect four except the A3SS differential splicing event of gene TMX2, while Cuffdiff only detected TMX2 as a differentially spliced gene.

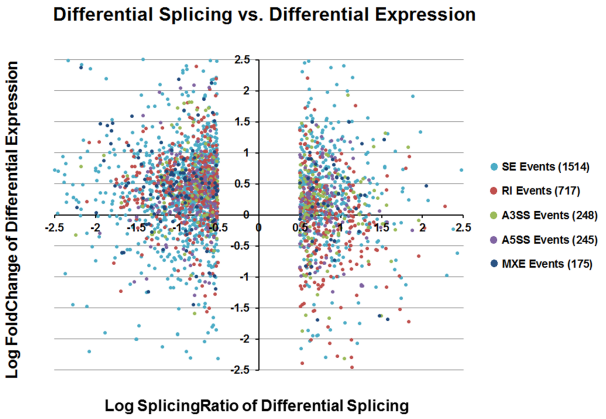

Figure 3. The plot of differential splicing vs. differential expression (H1 vs. H1-npc data set).

c) Comparison to MATS and Cufflinks results: We also compared the results of the detected differentially spliced genes by dSpliceType with MATS ([17]) and Cufflinks/Cuffdiff ([8]) for the H1 and H1-npc RNA-Seq data set. We ran MATS and Cufflinks/Cuffdiff with their default settings and considered a gene differentially spliced as same as the simulation studies. Please note, for dSpliceType and MATS, a gene may contain multiple splicing events.

As shown in Figure 9, dSpliceType predicted a total of 2,028 differentially spliced genes, while 1,169 and 837 genes were predicted by MATS and Cufflinks/Cuffdiff, respectively. The results show that dSpliceType can predict more differentially spliced genes than the other two. Moreover, there were amount of overlapped genes among the three methods, in which 149, 431 and 112 genes were predicted by all three methods, both

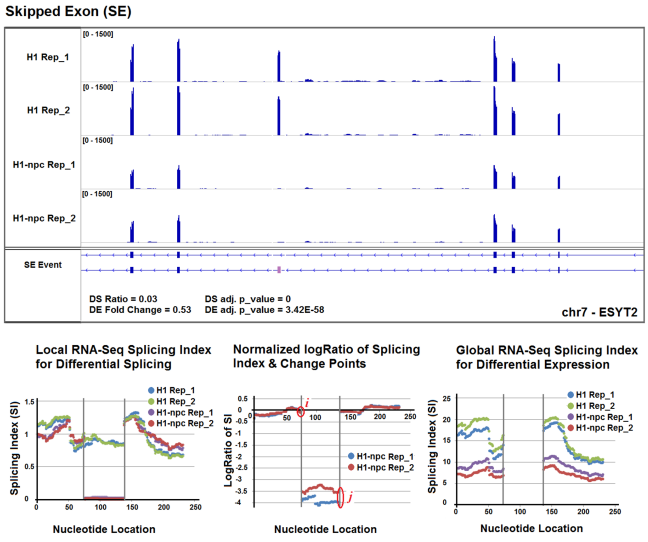

Figure 4. A skipped exon (SE) differential splicing event with down-regulated differential expression is detected for the gene chr7 - ESYT2 with two change points *i* and *j* at the ending locations of the left common exon and the spliced exon.

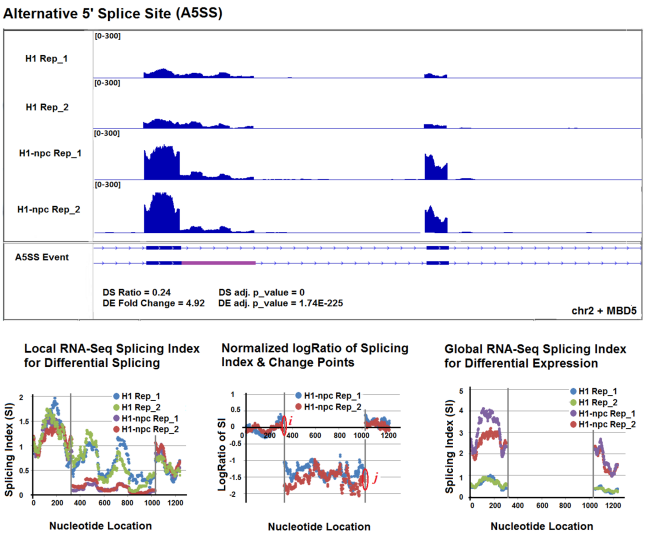

Figure 5. An alternative 5' splice site (A5SS) differential splicing event with up-regulated differential expression is detected for the gene chr2 + MBD5 with two change points *i* and *j* at the ending locations of the left common exon and the spliced exonic region.

MATS and dSpliceType only, and both Cufflinks/Cuffdiff and dSpliceType only, respectively. It shows that dSpliceType and MATS are more consistent, and Cufflinks/Cuffdiff predicts fewer differentially spliced genes. On the other hand, the results are likely to indicate that the three methods are complementary while dSpliceType can predict more, since amount of genes were predicted uniquely by each of the methods with 1,336, 499 and 487 genes by dSpliceType, MATS and Cufflinks/Cuffdiff, respectively. In overall, the trend of the comparison results is similar to that of simulation studies.

2) Human autism disease data set

We also applied dSpliceType to a public Illumina Genome Analyer II single-end RNA-Seq data set of human autism spec-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

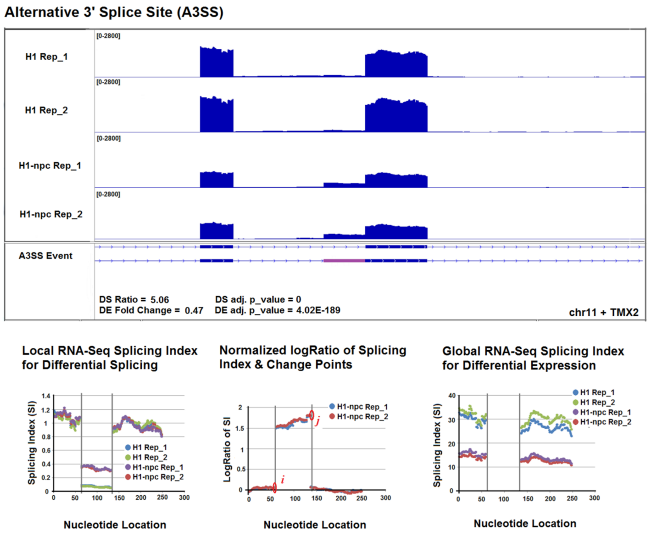

Figure 6. An alternative 3' splice site (A3SS) differential splicing event with down-regulated differential expression is detected for the gene chr11 + TMX2 with two change points i and j at the ending locations of the left common exon and the spliced exonic region.

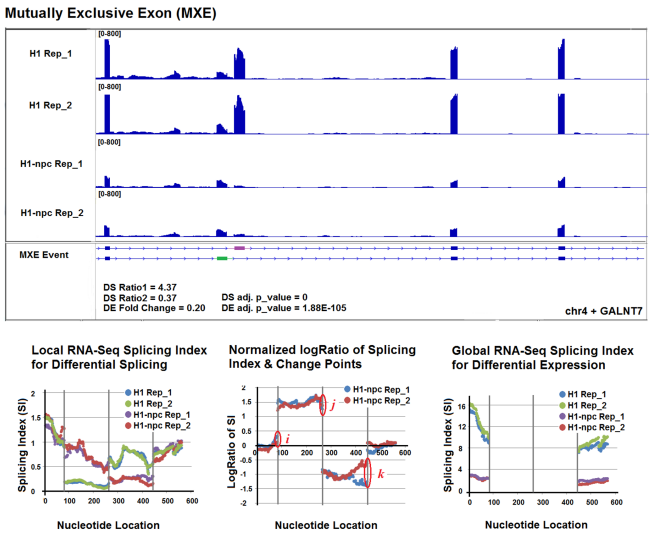

Figure 8. A mutually exclusive exon (MXE) differential splicing event with down-regulated differential expression is detected for the gene chr4 + GALNT7 with three change points i, j and k at the ending locations of the left common exon and the two spliced exons.

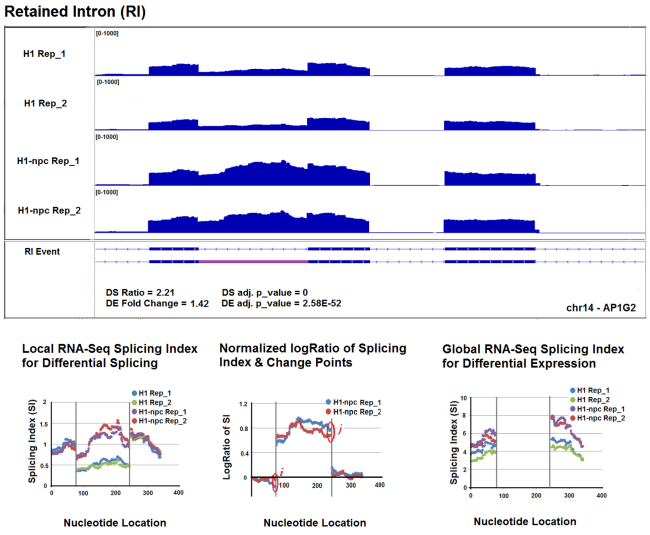

Figure 7. A retained intron (RI) differential splicing event with up-regulated differential expression is detected for the gene chr14 - AP1G2 with two change points i and j at the ending locations of the left common exon and the spliced exonic region.

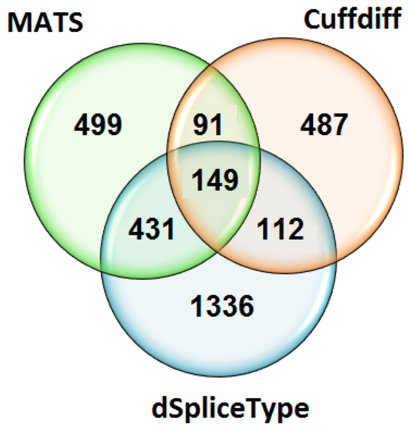

Figure 9. The comparison of the detected differentially spliced genes by dSpliceType, MATS and Cuffdiff (H1 vs. H1-npc data set).

trum disorder (ASD) disease [30]. The data set contains three control samples and three autism samples with down-regulated splicing factor A2BP1, which can be accessed from Gene Expression Omnibus (GEO) with accession number GSE30573. About 40 to 50 million short reads (74bp) were sequenced for each replicate. We aligned the RNA-Seq reads to the human reference genome and applied the multivariate dSpliceType [20] to detect the differential splicing events of human autism disease. By using a filter of $|\log \text{SplicingRatio}| > 0.5$, dSpliceType detected 243 genes with significant differential skipped exon splicing events. Among these, 72 differentially spliced genes had been reported in Voineagu *et al.* [30], demonstrating a substantial overlap between the two studies as shown in Figure 10. Figure 11 shows a SE event detected by dSpliceType that was not reported in Voineagu *et al.* [30]. The multivariate dSpliceType exclusively detects 171 differentially skipped exons that warranties further studies.

IV. Discussion and Conclusion

Compared to DNA microarrays, RNA-Seq holds a strong promise for dissecting and comparing transcriptomes at a higher resolution. In microarray data, the gene expression signals are captured in probe-level intensities while in RNA-seq data, they were captured in short read counts. After aligning these reads to a reference genome or *de novo* assembly, two formats of gene expression data from RNA-seq can be used: read counts and read coverage. Read counts based methods usually require sequencing a certain amount of reads or a certain depth to ensure the computational methods work properly. Therefore, these methods are not particularly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

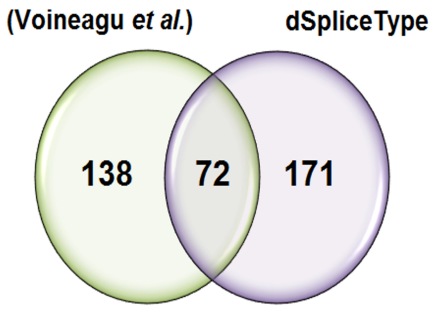

Figure 10. The comparison of the differentially spliced genes with Skipped-Exon (SE) events detected by the multivariate dSpliceType and those reported in Voineagu *et al*. [30].

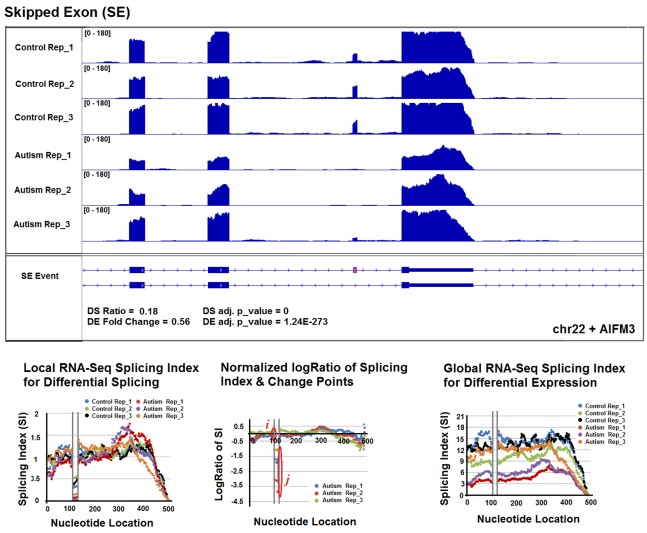

Figure 11. A Skipped Exon (SE) differential splicing event with down-regulated differential expression is detected for the gene chr22 + AIFM3 with two change points i and j at the ending locations of the left common exon and the spliced exon.

effective in analyzing gene expression with low abundances. However, read-coverage based methods overcome the limitation in that the data is available for each nucleotide albeit subject to moderate variability. The generalized dSpliceType is designed to utilize base-wise read coverage signals to detect both differential expression and differential splicing events with a high sensitivity.

In detecting differential splicing events, the goal is to sequentially detect the abrupt changes along the transcript sequence. Instead of using complex model for bias correction, we used ratio of normalized RNA-Seq splicing indexes between conditions to eliminate sequencing biases from RNA-Seq. Thus, the sharpening signal changes on the exon boundaries of splicing events can be easily identified as change points, even if the read coverage is relatively low. We employed a multivariate conditional normal model to capture the sequential dependency of the read coverage signals, and detect differential splicing events by comparing SIC scores between models with or without change points. Model parameters can be estimated accurately due to a large number of nucleotides, e.g. exceeds 100, on a common or spliced exonic region.

In detecting differential expression, the goal is to compare the mean gene expression between case and control conditions. The large number of read coverage signals permit the variance of the mean gene expression in each condition to be estimated more accurately and robustly. For clean data with high correlation existing between replicates, we used a univariate statistical test to detect and prioritize differential expression according to statistical significance. Otherwise, we use a multivariate normal model to detect and prioritize differential expression by calculating a Hotelling multivariate statistic and associated p-values.

The massive increase in RNA-Seq data brought in more opportunities at the same time created unprecedent computational changes. The generalized dSpliceType tool is built on parametric models therefore model parameters are available in a closed solution. Thus it is very computational efficient for big RNA-seq data. Moreover, the computational complexity of converting read alignment results to read coverage signals is linear with regard to the amount of read counts. Thus, dSpliceType is emerged as a scalable tool for gene expression analysis using RNA-seq. As one-of-the-kind read coverage based method, we believe that the generalized dSpliceType can be applied to RNA-Seq data from multiple sequencing platforms regardless of read length. dSpliceType is expected to be more powerful with the ever-increasing sequencing coverage depth.

REFERENCES

[1] H. Li, D. Zhu, and M. Cook, "A statistical framework for consolidating "sibling" probe sets for affymetrix genechip data," *BMC Genomics*, vol. 9, no. 1, p. 188, 2008.

[2] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116–5121, Apr. 2001.

[3] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *STAT. APPL. GENET. MOL. BIOL*, vol. 3, no. 1, 2004.

[4] H. Keren, G. Lev-Maor, and G. Ast, "Alternative splicing and evolution: diversification, exon definition and function," *Nature Reviews Genetics*, vol. 11, no. 5, pp. 345–355, 2010.

[5] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.

[6] T. A. Cooper, L. Wan, and G. Dreyfuss, "RNA and disease," *Cell*, vol. 136, no. 4, pp. 777–793, 2009.

[7] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.

[8] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, 2012.

[9] N. Deng, C. G. Sanchez, J. A. Lasky, and D. Zhu, "Detecting splicing variants in idiopathic pulmonary fibrosis from non-differentially expressed genes," *PloS One*, vol. 8, no. 7, p. e68352, 2013.

[10] M. Gonzàlez-Porta, M. Calvo, M. Sammeth, and R. Guigó, "Estimation of alternative splicing variability in human populations," *Genome research*, vol. 22, no. 3, pp. 528–538, 2012.

[11] N. Deng, A. Puetter, K. Zhang, K. Johnson, Z. Zhao, C. Taylor, E. K. Flemington, and D. Zhu, "Isoform-level microRNA-155 target prediction using RNA-seq.," *Nucleic Acids Research*, vol. 39, no. 9, 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2015.2388593, IEEE Transactions on NanoBioscience

[12] M. Aschoff, A. Hotz-Wagenblatt, K.-H. Glatting, M. Fischer, R. Eils, and R. König, "Splicingcompass: differential splicing detection using RNA-Seq data," *Bioinformatics*, vol. 29, no. 9, pp. 1141–1148, 2013.
[13] D. Singh, C. F. Orellana, Y. Hu, C. D. Jones, Y. Liu, D. Y. Chiang, J. Liu, and J. F. Prins, "FDM: a graph-based statistical method to detect differential transcription using RNA-seq data," *Bioinformatics*, vol. 27, no. 19, pp. 2633–2640, 2011.
[14] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Research*, vol. 22, no. 10, pp. 2008–2017, 2012.
[15] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nature Methods*, vol. 7, no. 12, pp. 1009–1015, 2010.
[16] J. Wu, M. Akerman, S. Sun, W. R. McCombie, A. R. Krainer, and M. Q. Zhang, "SpliceTrap: a method to quantify alternative splicing under single cellular conditions," *Bioinformatics*, vol. 27, no. 21, pp. 3010–3016, 2011.
[17] S. Shen, J. W. Park, J. Huang, K. A. Dittmar, Z.-x. Lu, Q. Zhou, R. P. Carstens, and Y. Xing, "MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data," *Nucleic Acids Research*, vol. 40, no. 8, pp. e61–e61, 2012.
[18] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, A. Monroy, P.-F. Kuan, S. M. Hammond, L. Makowski, *et al.*, "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq," *Nucleic Acids Research*, vol. 41, no. 2, pp. e39–e39, 2013.
[19] N. Deng and D. Zhu, "Detecting various types of differential splicing events using RNA-Seq data," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, p. 124, ACM, 2013.
[20] N. Deng and D. Zhu, "dsplicetype: A multivariate model for detecting various types of differential splicing events using rna-seq," in *Bioinformatics Research and Applications* (M. Basu, Y. Pan, and J. Wang, eds.), vol. 8492 of *Lecture Notes in Computer Science*, pp. 322–333, Springer International Publishing, 2014.
[21] Y. Xing, P. Stoilov, K. Kapur, A. Han, H. Jiang, S. Shen, D. L. Black, and W. H. Wong, "MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays," *RNA*, vol. 14, no. 8, pp. 1470–1479, 2008.
[22] M. L. Eaton, *Multivariate statistics: a vector space approach*. Wiley New York, 1983.
[23] J. Chen, *Parametric statistical change point analysis*. Birkhauser Boston, 2012.
[24] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
[25] D. Zhu, Y. Li, and H. Li, "Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data," *Bioinformatics*, vol. 23, pp. 2298–2305, Sept. 2007.
[26] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigó, and M. Sammeth, "Modelling and simulating generic RNA-Seq experiments with the flux simulator," *Nucleic acids research*, vol. 40, no. 20, pp. 10073–10083, 2012.
[27] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome biology*, vol. 14, no. 4, p. R36, 2013.
[28] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.
[29] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.
[30] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

Dongxiao Zhu is currently an Assistant Professor at Department of Computer Science, Wayne State University. From 2008 to 2011, he was an Assistant Professor at Department of Computer Science, University of New Orleans. From 2006 to 2008, he worked at Stowers Institute for Medical Research as a Biostatistician. He received his Ph.D. from University of Michigan in 2006. His research interests have been in areas of computational biology, bioinformatics, health informatics and the interface with data mining, machine learning and pattern recognition. Dr. Zhu has published over 40 peer-reviewed publications and 7 book chapters and he served on 6 editorial boards of bioinformatics journals. Dr. Zhu's research has been supported by National Institutes of Health (NIH), National Science Foundation (NSF), State of Louisiana and private agencies and he has served on multiple NIH and NSF grant review panels. Dr. Zhu has advised numerous students at undergraduate, graduate and postdoctoral levels.

Nan Deng is a currently Senior Research Bioinformatican at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center. She received her Ph.D. in Computer Science from Wayne State University in 2014. Dr. Deng has extensive experiences in algorithm design and implementation, statistical model building and next generation sequencing (NGS) data analysis. She has developed novel algorithms and computational software tools for transcriptome quantification, characterization and identification using RNA-Seq, in particular detecting various types of differential splicing events between healthy and diseased human transcriptomes. Her research interests have been extended to study life-threatening human diseases by integrating genomics, transcriptomics, epigenetics and proteomics.

Changxin Bai received his B.S. degree in Electronic Engineering from XiDian University. He is currently a PhD candidate in Department of Computer Science at Wayne State University. His primary research interests lie in bioinformatics, health informatics and machine learning.