1-1-2014

Algorithms And Tools For Computational Analysis Of Human Transcriptome Using Rna-Seq

Nan Deng
Wayne State University,

ALGORITHMS AND TOOLS FOR COMPUTATIONAL ANALYSIS OF HUMAN TRANSCRIPTOME USING RNA-SEQ

by

NAN DENG

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: COMPUTER SCIENCE

Approved by:

_____

Advisor Date

_____

_____

_____

DEDICATION

To my MOTHER, FATHER and HUSBAND

ACKNOWLEDGEMENTS

First and most importantly, a tons of thanks are given to my advisor, Dr. Dongxiao Zhu, for bringing me into bioinformatics field, encouraging me to develop research ideas independently and providing me his consistent and endless support including financial support from Dr. Zhu's NIH grant R21LM010137. Undoubtedly, I have greatly benefited from his academic guidance and in-depth knowledge of bioinformatics. I would like to extend my special thanks to Dr. Joseph A. Lasky and Dr. Erik Flemington of Tulane University for their fruitful collaborations, as well as Dr. Xuewen Chen and Dr. Jing Hua who provided constructive suggestions in writing my dissertation.

I also thank my former bioinformatics group colleagues Dr. Guorong Xu, Dr. Lipi Acharya and Dr. Thair Judeh for their collaboration, support and help in my research. I would also like to thank the Department of Computer Science at Wayne State University for the generous funding and resources they have provided in supporting my research.

Finally, special thanks should be given to my family members, especially my beloved husband Zunguo Dai who has constantly taken care of my life, supported and encouraged me to excel in my research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Background and Introduction

1.1 Central Dogma of Molecular Biology

The central dogma of molecular biology describes the transfer of genetic sequence information within a living cell [20]. Accordingly, cell information usually is transcribed from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA). It is then translated from RNA to protein as shown in Figure 1.1. It involves two major processes: transcription and translation. In a cell nucleus, the transferring of a section of DNA sequence information to a pre-messenger RNA (mRNA) molecule is called transcription. The pre-mRNA is then further alternatively spliced to produce mature mRNA. Eventually, it is transported from the nucleus into the cytoplasm and translated by ribosomes to produce proteins. The process from mRNA to protein is called translation. This dissertation research is primarily focused on two novel algorithms/tools and a computational workflow for the analysis of human transcriptomes.

1.2 Genome and Transcriptome

The genome of an organism is its entire biological hereditary information that is essential for cell growth, replication and apoptosis [2]. Biological information is sequentially represented by four nucleotides: Adenine(A), Guanine(G), Cytosine(C) and Thymine(T). This information is stored in a double-stranded DNA sequence. In human cells, there are 23 pairs of chromosomes including 22 pairs of autosomes and a pair of sex chromosomes (XX for females and XY for males). Each chromosome contains a long DNA sequence that stores a proportion of hereditary information. Furthermore, each DNA molecule contains many genes represented by individual

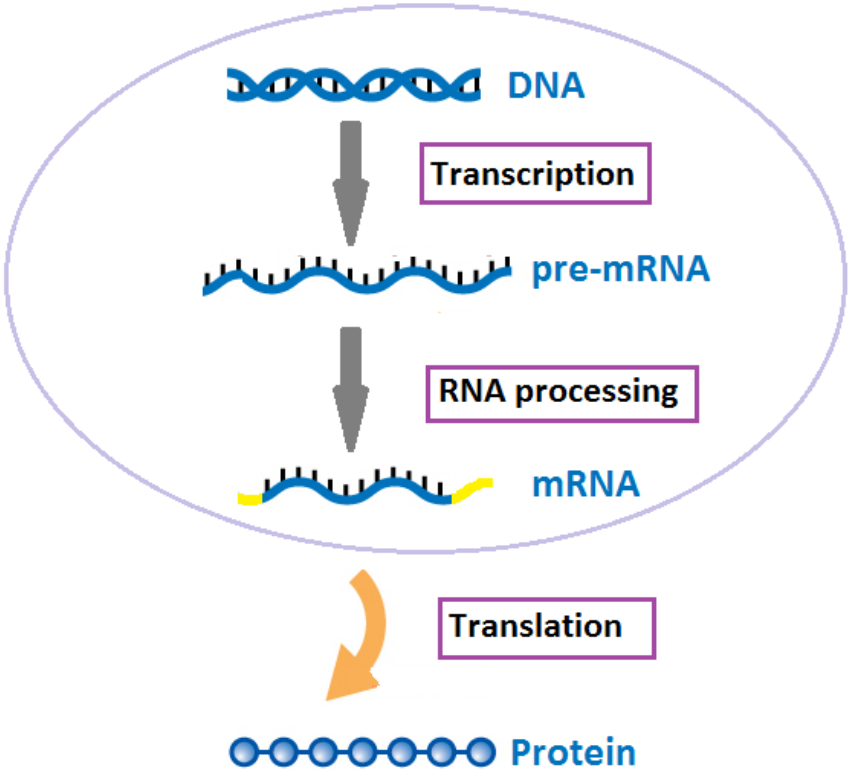


Figure 1.1: Central dogma of molecular biology [2].

segments of DNA. Genes encode genetic information that is necessary to produce RNA or protein molecules. More than 20,000 protein-coding human genes are considered in this dissertation.

An organism‘s transcriptome consists of an entire set of all transcripts, such as mRNAs, small RNAs and long intervening non-coding RNAs (lincRNAs) , in one or a population of cells under a particular development stage or physical condition [121]. In eukaryotic cells, one gene can produce more than one mRNA transcript by alternative splicing, which greatly increases the diversity and complexity of the transcriptome. Transcriptomes are known to vary corresponding to normal cellular

development and differentiation or a certain condition caused by environmental factors or diseases. Thus, studying transcriptomes may lead to a better understanding of cellular processes and progression of human diseases [17, 45, 70, 87, 117, 118].

1.3 Alternative Splicing

Genes consist of exons and introns. Exons are transcribed and spliced into mature mRNA for protein synthesis. Introns are non-coding DNA sequences between exons that do not code for proteins, and they are typically removed by splicing. Alternative splicing is a gene regulation mechanism in which different exons can be combined together [2]. As shown in Figure 1.2, a pre-mRNA can be processed to produce several mature mRNAs by removing introns and concatenating different exons via alternative splicing [52, 118]. The mature mRNAs are further translated into protein isoforms with different structures and functionalities [52, 118].

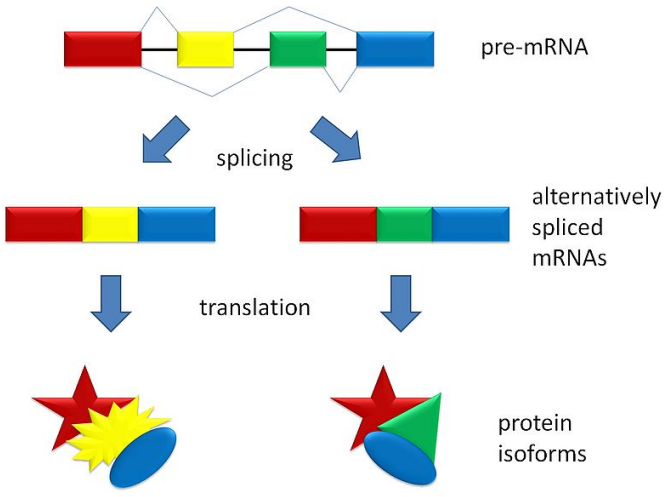


Figure 1.2: Alternative splicing (From Wikipedia).

Alternative splicing plays important roles in many biological processes including diseases [52, 119]. It markedly increases the diversity of transcriptome and proteome by producing multiple mRNAs and proteins from the same gene with inclusion

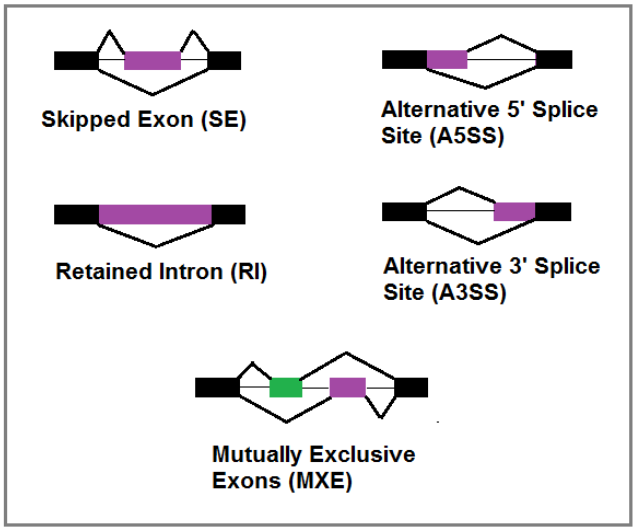


Figure 1.3: Five types of alternative splicing events [28].

and exclusion of specific exons. More than 90% of multiexonic human genes are alternatively spliced [87]. As a result, the total number of transcripts or proteins are much more than the number of protein-coding genes. Several types of alternative splicing events have been observed in biological experiments [28, 118], and Figure 1.3 demonstrates the five well-known types of alternative splicing events, including skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A5SS), retained intron (RI) and mutually exclusive exons (MXE).

Due to the limitations of earlier sequencing technologies, such as DNA microarray, the human transcriptomes in different tissues or biological conditions can not be fully explored and compared. However, examining the differences between human transcriptomes may help lead to a better understanding of biological processes and yield new insights into diseased cell development and differentiation as well as new hypotheses on potential biomarkers for human diseases [17, 118, 119].

The emergence of the next-generation sequencing technologies, in particular, RNA-Seq, has provided unprecedented opportunities to interrogate the entire tran-

scriptome, including detecting differences of pre-mRNA alternative splicing between human transcriptomes [87, 121]. As studies increasingly shift from DNA microarray to RNA-Seq, it holds the promise for a better characterization and biological understanding of transcriptomes [23, 87, 121].

1.4 Next-generation Sequencing

Over the past decade, first-generation sequencing such as Sanger sequencing and microarray technologies have dominated in the areas of genome and transcriptome analyses, respectively. Although these technologies have proved useful, there is still an urgent need for new technologies to sequence large amounts of human genomes and transcriptomes [76]. Recent sequencing techniques are termed as next-generation sequencing (NGS) or high-throughput sequencing technologies. Comparing with the low-throughput and long sequencing of the Sanger technique, the massively parallel and short read sequencing provided by NGS technologies can complete genome projects in weeks while taking several years using the higher cost Sanger technology [76]. Since 2005, the advent of NGS technologies has not only expanded our horizons, but also changed our ways of thinking and conducting biomedical research [76]. One major advantage of NGS is the capability of producing enormous amount of sequencing data at lower costs. Millions of short reads can be directly sequenced from DNA or RNA molecules at nucleotide base-pair resolution. This feature provides more opportunities to interrogate whole genomes or transcriptomes but not limited to just determining the order of ATCG sequences or the expression abundance of annotated genes. The ability of NGS to sequence whole human genomes and transcriptomes to a great depth makes it possible to enhance our understanding of how genomic and transcriptomic differences affect our health. Also, large-scale comparative and evolu-

tionary studies can be performed among many homologous organisms, which was not possible before the emergence of NGS sequencing technologies [76].

Due to the variety of NGS applications, different types of NGS data are generated, such as DNA-Seq, RNA-Seq, ChIP-Seq, microRNA-Seq and so on. Several NGS platforms coexist to support genome and/or transcriptome analyses with differences in read length, cost and run time. For instance, the Illumina/Solexa Genome Analyzer II and HiSeq platforms have been used to analyze mouse and human transcriptomes [22, 23, 78]. Applied Biosystems' Solid Sequencing (ABI SOLiD) has been applied to profile transcriptome of mouse embryonic stem cells and whole-genome mutation of yeast [16, 104]. Research on detection of SNPs from the maize transcriptome and determination of microbal diversity has been conducted using Roche/Life Sciences' 454 Sequencing [6, 88]. Single-molecule sequencing technologies, including Helicos BioSciences, Pacific Biosciences and Oxford Nanopore Technologies, have recently entered the market and may in the near future impact genomics research [42]. This dissertation is mainly focused on transcriptome analysis using RNA-Seq data.

1.5 RNA-Seq Technology

RNA-Seq is a revolutionary tool for profiling transcriptomes by sequencing mRNAs of a sample using high-throughput sequencing technologies [121]. It has rapidly become a promising approach to study eukaryotic transcriptomes. Prior to RNA-Seq, hybridization-based (e.g., DNA microarrays [97]), sequence-based (e.g., Sanger sequencing [35] and expressed sequence tags (ESTs) [1]) or tag-based (e.g., serial analysis of gene expression (SAGE) [116], cap analysis of gene expression (CAGE) [55] and massively parallel signature sequencing (MPSS) [11]) approaches have been used for years to quantify and decipher transcriptomes. However, These methods have

different technical limitations [121]. Microarrays are limited by existing knowledge of genome sequences, cross-hybridization and saturation signals. Sequence-based methods are expensive and the sequencing throughputs are not high. Tag-based methods are not ideal because of a large portion of very short tags from the technology. These traditional sequencing technologies have prevented researchers from better interrogating transcriptomes [121].

Recently, the advent of RNA-Seq has dramatically changed the way we study transcriptomes and has been applied to different organisms [16, 77, 78, 80, 122]. By sequencing at nucleotide resolution of millions of short reads directly from mRNA molecules, RNA-Seq has a number of applications beyond those of existing array techniques. These include genome annotation [26], quantification of relative transcript abundances, identification of differentially expressed transcripts [113], discovery of novel transcript isoforms of genes [109], comprehensive identification of gene/transcript fusion in cancer [73], and transcriptome assembly [38]. For many biological applications, microarrays have been replaced [121] by RNA-Seq as it continually becomes cheaper. RNA-Seq is capable of identifying and quantifying both annotated and novel transcripts with providing the information of both exonic and exon-exon junction reads. Thus, it allows researchers to detect and quantify differences of transcriptomes more precisely regarding alternative splicing [76, 121]. In addition, DNA microarrays provide a relatively small dynamic range of gene expression because its sensitivity is inherently limited by signal saturation, non-specific hybridization and probe design while RNA-Seq does not have this limitation [121]. Also, RNA-Seq has low background signal since most of the reads can be uniquely mapped to the reference genome [121].

Figure 1.4 shows a typical RNA-Seq sequencing procedure followed by read alignment [121]. A population of mRNA molecules with poly(A) tails is first converted

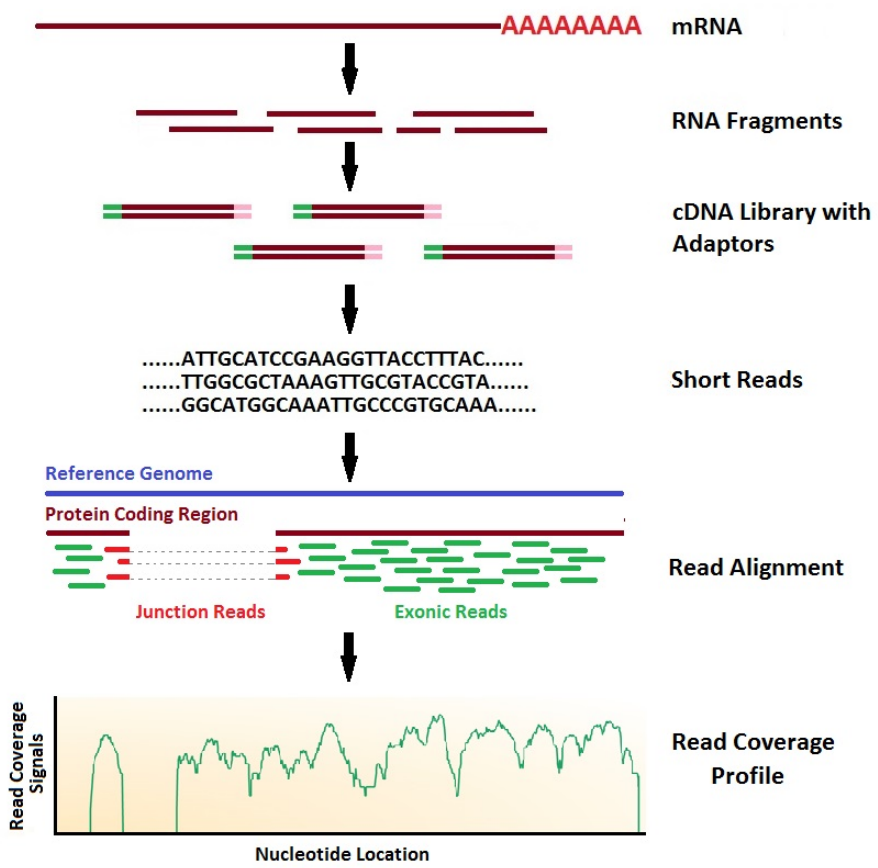


Figure 1.4: A typical RNA-Seq sequencing procedure, alignment and read coverage profile [121].

into a library of cDNA fragments via either RNA or DNA fragmentation. Second, sequencing adaptors are attached to each cDNA fragment on both ends. Depending on sequencing platforms, amplification procedure is optional. Finally, cDNA fragments are sequenced by a high-throughput sequencer to generate short reads from either one end (single-end reads) or both ends (paired-end reads). To analyze RNA-Seq data, these short reads are usually aligned to a reference genome and followed by the downstream analysis. If the reference genome is not available or is of low quality, transcriptome *de novo* assembly may be applied.

1.6 Short Read Alignment

Once high-quality RNA-Seq reads have been generated, the first step of a typical data analysis is to map these short reads back to a reference genome. Since mRNAs concatenate exons by removing introns, the RNA-Seq reads may be originated from either exonic regions or exon-exon junction regions as shown in Figure 1.5a. In the reference genome, exons are separated by introns, thus exonic reads can be fully mapped to the regions where they originated. On the other hand, junction reads would be aligned in two parts spanning the ends of two concatenated exons (Figure 1.5b). There are several short read aligners. Bowtie [60] and Bowtie2 [59] use algorithms based on the Burrows-Wheeler transform [12] and FM-index [33]. Novoalign (http://www.novocraft.com/) is based on the Needleman-Wunsch algorithm [81]. Burrows-Wheeler Aligner (BWA) is based on the Burrows-Wheeler transform and Smith-Waterman [105] genome alignment methods. There are also other existing methods for sequence alignment, such as [66, 68, 103]. For junction reads, Tophat [112] and Tophat2 [53] are commonly used splice junction mappers.

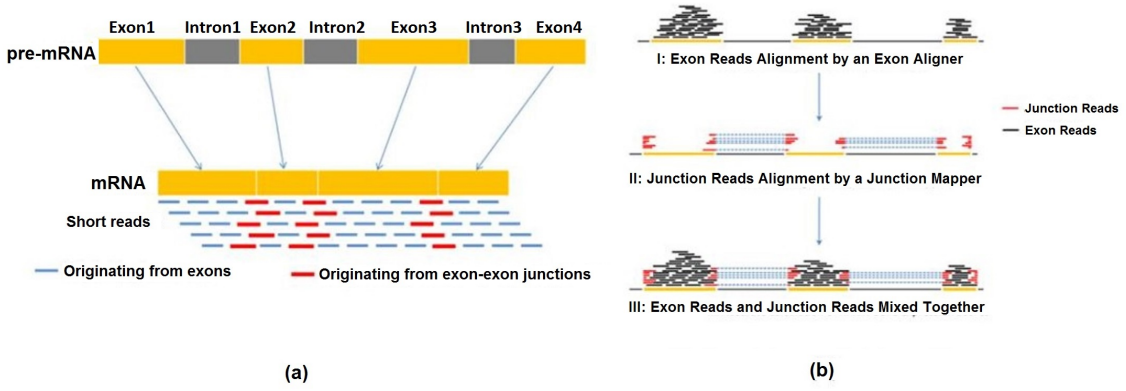


Figure 1.5: The illustration of exonic reads and junction reads [128]. (a) Short reads originating either from exons or from exon-exon junctions. (b) The illustration of the aligned exon reads and junction reads.

1.7 Common Format of Alignment Files

The read alignment result, including both single-end and paired-end reads, is typically stored in Sequence Alignment/Map (SAM) files [65], in which the original coordinates of the aligned reads are recorded. SAM is a Tab-delimited text format. It is easy to read and understand; however, the size of a SAM file is often up to several gigabytes, which may lead to storage challenges. Therefore, the binary SAM (BAM) format [65] is commonly used to store compressed alignment result and is then parsed by NGS-related software. Besides SAM/BAM alignment files, BED files of junction reads mapped using Tophat [112] or Tophat2 [53] are commonly used to discover novel transcripts and to explore splicing patterns. Also, WIG and BEDGRAPH files are used to store the information of numbers of covered reads at each base (nucleotide) location and are therefore useful for visualization and utilization of read coverage signals.

1.8 Review of Transcriptome Quantification Method using RNA-Seq

As RNA-Seq is quantitative, quantifying expression abundance at either gene-level or isoform-level via measurement of mRNA expression level is considered more accurate than microarrays [121]. Transcriptome quantification is a fundamental and important process for many bioinformatics applications. As studies increasingly shift from DNA microarrays to RNA-Seq, the latter approach holds great promise for transcriptome quantification. Instead of absolute transcript abundances, RNA-Seq allows estimating relative transcript abundances since mRNA molecules are proportionally sampled from experiments [86].

1.8.1 Estimation of Gene-level Expression Abundance

One of the first methods for RNA-Seq transcriptome quantification uses read counts information from alignment results [78]. The model takes uniquely mapped single-end reads to a gene as input and outputs the gene expression abundance. This approach calculates gene-level expression abundance under two assumptions: 1) reads are uniformly sampled among all positions from each transcript, and 2) the number of reads from a given region follows a Possion distribution. According to [78], gene expression abundance can be measured by the number of mapped Reads Per Kilobase of exon model per Million mapped reads (RPKM). RPKM is calculated as the number of reads that fall into exonic regions of a gene normalized by the total length of exons and the total number of uniquely mapped reads in the RNA-Seq experiment as shown in the following equation:

$$\text{RPKM} = 10^9 \times \frac{C}{L \times N}, \tag{1}$$

where C is the total number of reads mapped in the exonic regions of a gene, L is the total length of exons and N is the total number of uniquely mapped reads in the sequencing run. The value of RPKM is a normalized number of read counts and can thus be used to detect differential expression during cell development and between two conditions.

1.8.2 Estimation of Isoform-level Expression Abundance

With the development of transcriptome quantification methods, other models have been developed for genes with multiple transcripts. In this case, reads sequenced

from one transcript may be mapped to more than one transcript, because the exon where the reads are mapped to may be shared by multiple transcripts within a gene. Due to the ambiguity of read mapping and the complexity of transcript structures, different types of models have been developed for estimating relative abundances of multiple transcripts. Most of these models can be categorized into statistical models or mathematical models.

Statistical models

In [50], a Poisson distribution is used to model the number of reads mapped to a given region. It formulates the transcript quantification as an optimization problem by maximizing a likelihood function. In [92] (software: Solas), Poisson and multinomial distributions are employed to model the number of reads mapped to each transcript and the number of reads mapped to each exon of a transcript, respectively. It uses an EM algorithm to estimate the proportion of each isoform. In [31] (software: IsoInfer), the authors employ a normal distribution to approximately model the read counts in each exon. Transcript quantification is then formulated as an optimization problem solved by quadratic programming.

In contrast to modeling the read counts in a given genomic region, [64] (software: RSEM), [114] (software: Cufflinks) and [22] (software: SAMMate(RAEM)) directly use the information of read originating positions to build a mixture model. This is then followed by an EM algorithm to find the maximum likelihood estimation of the mixture proportions of isoforms.

Mathematical models

Besides statistical models, mathematical methods using a constrained least square attempt to minimize the differences between the observed and the expected

read coverage signals in terms of estimated parameters corresponding to relative transcript abundances [9] (software: rQuant.web) and [82, 83] (software: SAMMate (SASeq)).

1.8.3 Fragment Length Distribution for Paired-end Reads

In addition to developing transcriptome quantification methods for single-end reads, methods have been developed to accommodate paired-end reads sequenced from the ends of the cDNA fragments. For paired-end reads, relative transcript abundances can be measured using Fragment Per Kilobase of transcript per Million mapped reads (FPKM), which is an extension of RPKM. These methods assign each fragment to its compatible transcripts according to the probability of the calculated effective fragment length based on the fragment length distribution. The first batch of paired-end transcript quantification methods include [114] (software: Cufflinks), [51] (software: MISO), [84] (software: IsoEM), [95], [63] (software: RSEM) and [32] (software: IsoInfer).

1.8.4 Bias Correction

Most of the aforementioned methods assume that reads or fragments are uniformly sequenced from each position of each transcript. However, positional [9] and sequencing [44, 106] biases have been discovered by some recent studies, and the biases can cause over- or under-estimation of expression levels. In [67], it was suggested to use variable rates for different positions to model read counts along each transcript. Two bias models were provided to predict the sequencing preference for each base location according to its surrounding bases. Other methods have modified the existing methods by adding bias correction. For example, in [124] the authors correct

positional bias on the model of [50], and [93] modifies the model of [114] by correcting both positional and sequencing biases. It also estimates simultaneously the bias and abundance parameters by maximizing the likelihood function. The modified methods have shown improved accuracy in estimating relative transcript abundances.

1.9 Review of Differential Splicing Detection using RNA-Seq

Estimation of transcript abundance has enabled the detection of differential splicing at the genome scale. In general, differential splicing refers to the difference in the relative abundances of isoform transcripts in a gene across samples from two conditions [46], such as healthy or diseased human transcriptomes, stages in cell development and differentiation, and different tissue types, e.g., brain and liver. Intuitively, differentially spliced genes can be detected by quantifying the discrepancy of estimated relative proportions of transcripts within a gene between two conditions. This type of methods estimates relative transcript abundances or proportions followed by a statistical test of relative proportions between two conditions to quantify the difference or discrepancy of isoform proportions. For example, Cufflinks/Cuffdiff [113] uses Jensen-Shannon divergence, the authors of [23] employ Pearson's Chi-square test of independence between conditions and isoforms using pseudo counts, and Hellinger distance is used by [37]. These approaches are powerful in directly detecting the change of isoform proportions. However, they rely on accurate estimation of transcript abundances, which is a challenging problem itself because of unknown positional and sequence-specific biases [9, 44, 106], unknown transcripts, the number of expressed transcripts, and the structures of transcripts among others.

The second type of methods detects differentially spliced genes by comparing read counts either on all exons within a gene, such as SplicingCompass [5], FDM [102]

and MMD [107], or differential usage of a single exon, e.g., DEXSeq [4]. This type of approaches can potentially detect differentially spliced genes but may not be able to specify the spliced regions and/or the associated types of alternative splicing.

The third type is event-based methods. Instead of detecting differences of individual transcripts or exon(s), this type of methods identifies differences in utilization of a skipped exon by isoform transcripts, such as ALEXA-seq [40], MISO [51], SpliceTrap [123] and MATS [100]. These methods focus on the detection of differentially skipped exon splicing event but are not designed for other types of alternative splicing events. A recently published method, DiffSplice [46], detects differential splicing events on read-alignment-based "alternative splicing modules". This approach does not rely on annotation databases. However, this method needs to estimate relative abundances of "alternative splicing module" paths [46] based on the assumption that reads are sequenced independently and uniformly from expressed transcripts [50]. Even though DiffSplice [46] and the updated version of MATS [100] are able to detect multiple types of splicing events, the computational cost of these two methods is relatively high.

1.10 Motivation

The high-throughput RNA-Seq technology provides unprecedented opportunities to study transcriptomes for a better understanding of transcriptional regulation and gene functionality in both normal cell development and progression of various human diseases. With millions of short reads, one of the most powerful advantages of RNA-Seq is its capability of capturing transcriptome dynamics across different tissues or conditions at the transcript isoform level [121]. However, due to the size and complexity of RNA-Seq data, typical problems faced by biomedical researchers are how

to extract information and gain biological insights from tremendous amounts of data. To aid biomedical researchers in studying and understanding transcriptomes, efficient computational tools for analyzing RNA-Seq data to detect differences in human transcriptomes, in particular the dynamics in terms of splicing patterns of transcript isoforms from all genes between healthy and diseased conditions, are urgently needed.

We have developed two algorithms and tools and a computational workflow using RNA-Seq to analysis human transcriptomes between healthy and diseased conditions. The first algorithm and tool is based on read count. It estimates relative transcript abundances using an EM algorithm. The second is based on read coverage signals. It utilizes sequential dependency of normalized base-wise read coverage signals and a change-point analysis followed by a parametric statistical hypothesis test using Schwarz Information Criterion (SIC) to detect significant differential splicing events in the form of five well-known types, including skipped exon (SE), retained intron (RI), alternative 3' or 5' splice sites (A3SS or A5SS), and mutually exclusive exons (MXE). Finally, a novel computational workflow is developed to jointly study human genes with differential expression and differential splicing.

1.11 Overview

This dissertation is organized into 5 chapters. Chapter 1 consists of a brief introduction of background information and motivation behind this work. Chapter 2 describes an EM-based algorithm and tool, Read Assignment Expectation Maximization (RAEM), for estimating relative transcript isoform proportion/abundance using RNA-Seq data. An application of this algorithm and tool to predict isoform-level microRNA-155 targets is also presented in this chapter. Chapter 3 presents a novel algorithm and tool, detection via Splicing Type (dSpliceType), to detect differential

splicing events between two conditions based on five well-known types of alternative splicing. The computational tool includes a univariate algorithm for comparing without replicates and a multivariate algorithm for comparing with replicates. Chapter 4 presents a joint RNA-Seq computational workflow of combining differential expression and differential splicing to dissect human diseases. We employed the workflow to detect differentially spliced genes without differential expression from a human lung disease, idiopathic pulmonary fibrosis (IPF). Finally, Chapter 5 concludes the research of this dissertation and gives several possible directions for future work. It should be noted that the work presented throughout this dissertation is largely based on and derived from original author contributions in [22, 23, 24, 25].

Chapter 2: RAEM : An Expectation-Maximization Algorithm and Tool to Estimate Relative Transcript Abundances using RNA-Seq [1]

Recently, more and more studies have switched from Microarray to RNA-Seq technologies making RNA-Seq the better choice for transcriptome analysis. In addition to gene expression, RNA-Seq also provides an opportunity to estimate relative transcript abundances more accurately [86]. Since mRNA molecules are sampled proportionally to absolute transcript abundances, accurate quantification of relative transcript abundances from the sequenced short reads is typically the first step for multiple RNA-Seq applications. Nevertheless, computational challenges remain in the problem of transcript quantification. As shown in Figure 2.1, spliced transcripts from a multiexonic gene are highly overlapped, and RNA-Seq reads are only sequenced from a small region of the entire set of protein-coding transcripts. After aligning short reads back to the reference genome, it is usually difficult to determine which transcript they originated from.

To overcome this challenge, we developed an EM-based algorithm and tool, Read Assignment via Expectation Maximization (RAEM), to solve the transcript quantification problem. RAEM estimates maximum likelihood proportions of transcripts within a gene. The details of the algorithm are described in Subsection 2.1.2. We also applied RAEM to a real RNA-Seq data set to predict isoform-level microRNA-155 targets.

[1] The content in this chapter is largely derived from original author text and contributions found in [22].

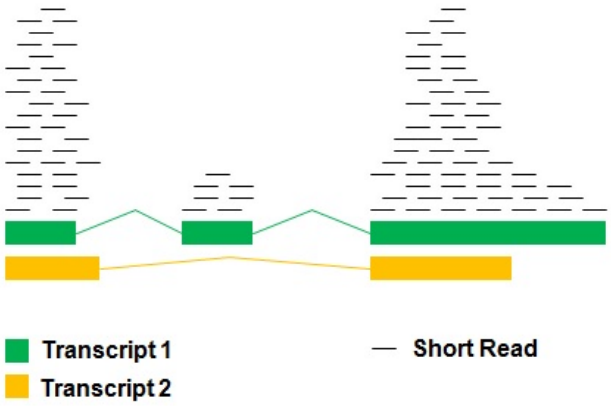


Figure 2.1: An illustration of short read pileup of transcript quantification.

2.1 Transcript Quantification

2.1.1 Expectation-maximization (EM) Algorithm

Expectation-maximization (EM) algorithm [21] is an iterative procedure to find maximum likelihood estimates (MLE) of parameters in statistical models of incomplete data problems. Typically, these models depend on unobserved latent variables. Finding a maximum likelihood solution of these models can not be solved in a closed-form, thus, the EM algorithm alternates between an expectation (E) step and a maximization (M) step until it converges.

Given a model consisting of a set of observed (or incomplete) data $\mathbf{Y}$, a set of unobserved data $\mathbf{Z}$ and a vector of unknown parameters $\mathbf{\Psi}$, along with a likelihood function $L(\mathbf{\Psi}; \mathbf{Y}, \mathbf{Z}) = p(\mathbf{Y}, \mathbf{Z}|\mathbf{\Psi})$, the MLE of the unknown parameters is determined by the marginal likelihood of the observed data

$$L(\mathbf{\Psi}; \mathbf{Y}) = p(\mathbf{Y}|\mathbf{\Psi}) = \sum_{\mathbf{Z}} p(\mathbf{Y}, \mathbf{Z}|\mathbf{\Psi}).$$

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following E and M steps:

E step:

$$Q(\mathbf{\Psi}|\mathbf{\Psi}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{Y},\mathbf{\Psi}^{(t)}}[\log L(\mathbf{\Psi}; \mathbf{Y}, \mathbf{Z})],$$

where the algorithm calculates the expectation of the log likelihood function over the conditional distribution of unobserved latent data $\mathbf{Z}$ given the observed data $\mathbf{Y}$ under the current estimates of the parameters $\mathbf{\Psi}^{(t)}$.

M step:

$$\mathbf{\Psi}^{(t+1)} = \arg\max_{\mathbf{\Psi}} Q(\mathbf{\Psi}|\mathbf{\Psi}^{(t)}),$$

where the algorithm updates the parameters by maximizing the Q function.

2.1.2 Transcript Quantification using EM Algorithm

In our study, suppose for each gene there are J annotated isoforms (denoted as $I_1, I_2, .., I_j$). For each short read we observed, p_j is used to denote the probability that this short read is generated from isoform I_j, where $j = 1, ..., J$ and $p_1 + p_2 + ... + p_j = 1$.

Suppose for one gene, we have N short reads (denoted as $R_1, R_2, ..., R_N$) and we know the correspondence between short reads and isoforms. Then we can use an $N \times J$ indicator matrix $Z = (z_{ij})$, where $i = 1, ..., N, j = 1, ..., J$) to represent the correspondence between short reads and isoforms (the matrix Z in Figure 2.2). If the ith read is generated from isoform I_j, then $z_{ij} = 1$; $z_{ij} = 0$ otherwise. So, for the matrix Z, each row indicates one short read, and only one isoform (column) for this row with value equal to 1. If the matrix Z is our observed matrix, it is easy to calculate the isoform proportions. The probabilities $(p_j, j = 1, ..., J)$ can be used for isoform proportion estimation. Intuitively, the number of short reads which are

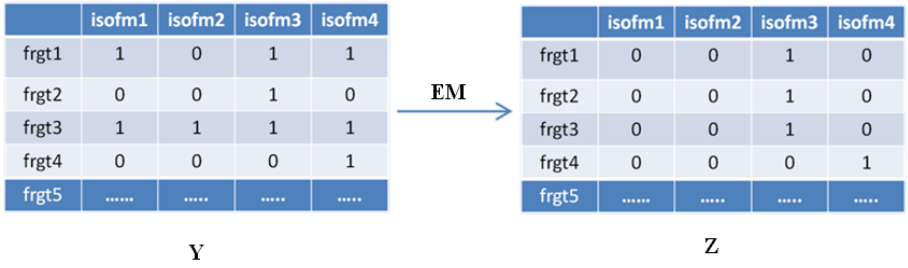


Figure 2.2: The illustration of RAEM algorithm [22].

generated from isoform I_j can be calculated by sum of jth column, correspondingly $n_j = \sum_{i=1}^{N} z_{ij}$, then the estimated isoform proportion $p_j = \frac{n_j}{N}$.

The read ambiguity issue in our study is that most of short reads are compatible with more than one isoform, shown as the matrix Y in Figure 2.2. Therefore, the indicator matrix Z is not fully observed. What we actually observed in a RNA-seq experiment is the indicator matrix $Y = (y_{ij})$ with $i = 1, ..., N, j = 1, ..., J$. Compare with the matrix Z, which has one and only one non-zero value in each row, the matrix Y has one or more than one non-zero value in each row. If $y_{ij} = 0$, then z_{ij} must be 0, but if $y_{ij} \neq 0$, then z_{ij} may or may not be 1. We define the indicator matrix Y as the observed cDNA fragment-compatible matrix, and the matrix Z as the unobserved cDNA fragment-originating matrix. We need to infer the matrix Z from the matrix Y.

Let's denote $P = (p_1, p_2, ..., p_j)$, where p_j is the mixture proportion for the isoform j. Given the observed cDNA fragment-compatible matrix Y, we estimate isoform proportions by finding the values of P that maximize the likelihood of the observed data:

$$L(P|Y) = \prod_{i=1}^{N} \sum_{j=1}^{J} p_j P(y_i|I = j).$$

The log-likelihood function is

$$\log L(P|Y) = \sum_{i=1}^{N} \log \sum_{j=1}^{J} p_j P(y_i|I=j).$$

The maximum likelihood estimates (MLE) of P can be written as $\arg\max_P L(P|Y)$, and we employed the EM algorithm to calculate the MLE of the isoform proportions $P = (p_1, p_2, ..., p_j)$ from our observed cDNA fragment-compatible matrix Y. The EM algorithm works in an iterative way, and it will be converged after numbers of iterations. Let's use $P^{(k)}$ to denote the isoform proportions computed after kth iteration. We initialized $P^{(0)} = (p_j^{(0)}, j = 1, ..., J)$ as $p_j^{(0)} = \frac{1}{J}$. Each iteration updates $P^{(k)}$ to $P^{(k+1)}$ through accomplishing the following E and M steps:

E-step:

$$\begin{aligned} z_{i,j}^{(k+1)} &= \mathrm{E}[z_{ij}|Y_i, P^{(k)}] = \Pr(z_{ij}=1|Y_i, P^{(k)}) \\ &= \frac{y_{i,j} p_j^{(k)}}{\sum_{j=1}^{J} y_{i,j} p_j^{(k)}}, \forall i, j, \end{aligned}$$

where $y_{ij} = \frac{1}{l_j}$ if the ith short read is compatible with isoform I_j, and $y_{ij} = 0$ otherwise. l_j is the length of isoform I_j. $\frac{1}{l_j}$ measures the probability of, given isoform I_j, the ith short read originated from any base location of isoform I_j, assuming that reads are uniformly sampled from each transcript.

M-step:

$$n_j^{(k+1)} = \sum_{i=1}^{N} z_{i,j}^{(k+1)}, \forall j,$$

$$p_j^{(k+1)} = \frac{n_j^{(k+1)}}{N}, \forall j,$$

The E-step updates the probabilities $z_{ij}^{(k+1)}$ that each short read generated from isoform I_j based on the current estimated isoform proportion set $P^{(k)}$, and M-step updates isoform proportion set from $P^{(k)}$ to $P^{(k+1)}$ based on $z_{ij}^{(k+1)}$. The EM algorithm iterates between E and M steps until convergence, i.e. $\sum_{j=1}^{J} \left| p_j^{(k+1)} - p_j^{(k)} \right| < \varepsilon$, where ε is an arbitrarily small positive number, i.e. 0.00001. To this end, we get the converged isoform proportion as $P^{(k+1)} = (p_j^{(k+1),j=1,\ldots,J})$. The transcript isoform expression abundance RPKM (for single-ended reads) or FPKM (paired-ended reads) can be calculated by the following equation :

$$\mathrm{RPKM_j/FPKM_j} = 10^9 \times \frac{p_j^{(k+1)} \times C}{l_j \times N},$$

where l_j is the sum of total exon length of isoform j in the gene, C is the total number of reads or fragments in exonic regions of the gene, and N is the total number of uniquely mapped reads or fragments in the sequencing run.

We named our EM-based transcript quantification method as Read Assignment Expectation Maximization (RAEM) and make note that the EM type algorithms have been used to solve multiple problems in bioinformatics. In particular, similar algorithms have been designed and applied to infer full-length isoforms using expressing sequence tags (ESTs) data [127] and RNA-seq data [64, 92]. We expect that the RNA-seq data works better with the EM type algorithm due to a much larger sample size and a much reduced number of compatible splicing isoforms.

Proof the concavity of the log-likelihood function

To guarantee that the EM algorithm is to reach a global maximum, we need to prove the concavity of the log-likelihood function of our model. The log-likelihood

function is:

$$\log L(P|Y) = \sum_{i=1}^{N} \log \sum_{j=1}^{J} p_j P(y_i|I=j).$$

Since the sum of concave functions is still a concave function, we only need to prove that

$$f(P) = \log \sum_{j=1}^{J} p_j P(y_i|I=j)$$

is concave. We denote $H(P)$ as the Hessian matrix of $f(P)$, and consider the (a,b)-th element of Hessian matrix $H(P)$ is:

$$\begin{aligned} H_{ab}(P) &= \frac{\partial^2 \log \sum_{j=1}^{J} p_j P(y_i|I=j)}{\partial p_a \partial p_b} \\ &= -\frac{P(y_i|I=a)P(y_i|I=b)}{(\sum_{j=1}^{J} p_j P(y_i|I=j))^2}. \end{aligned}$$

We can write $H(P)$ as $-d(P)x'x$, where $x = [P(y_i|I=1), ..., P(y_i|I=J)]$ is a vector and $-d(P) = \dfrac{1}{(\sum_{j=1}^{J} p_j P(y_i|I=j))^2}$ is a scalar. Because $-d(P) > 0$, and for any vector $y = [y_1, y_2, ..., y_J]$, we have

$$\begin{aligned} yH(P)y' &= y(-d(P)x'x)y' \\ &= -d(P)(yx')(yx')' \\ &= -d(P)(yx')^2 \\ &<= 0. \end{aligned}$$

Therefore, $H(P)$ is proved to be negative semidefinite, and both $f(P)$ and the log-likelihood function are concave. Given the concavity of the log-likelihood function, the local maximum of the EM algorithm is also the global maximum. Similar proof is given in [64, 50].

2.1.3 Simulation Studies

To assess the accuracy of RAEM in transcript quantification, we simulated RNA-seq experiments using FluxSimulator [39], a freely available software package that simulates whole transcriptome sequencing experiments with the Illumina Genome Analyzer. The software works by first randomly generating integer copies of each splicing transcript according to the annotation file provided by the user, followed by constructing an amplified, size-selected library and sequencing it in silico. The resulting cDNA fragments are then sampled uniformly at random for simulated sequencing, where the initial and terminal 25, 50 and 75 bp of each selected fragment are reported as reads. In our simulation studies, the human Ensembl ASTD database (version 57) was supplied to the software, along with the hg19 version of the human reference genome. In the ASTD annotation file, there are 100,297 protein coding transcripts, corresponding to 21,271 protein-coding genes. FluxSimulator then randomly assigned expression to 19,992 transcripts, corresponding to 10,343 genes. About 15-million single-end and 30-million paired-end RNA-seq 50-mer short reads were generated by size selection of fragments between 175 and 225 bases.

We applied RAEM to estimate the abundance of transcripts in each gene (Figure 2.3a for single-ended short reads, and Figure 2.3b for paired-end short reads). By using simulation data, both figures show that RAEM can estimate transcript abundance very accurately since the estimated abundances are highly correlated with the true abundances for the vast majority of the transcripts with high robust R^2. The very small portion of purple dots moving further away from the regression line (red line) correspond to those transcripts for which RAEM fails to estimate their abundance accurately in some situations, such as too many annotated transcripts

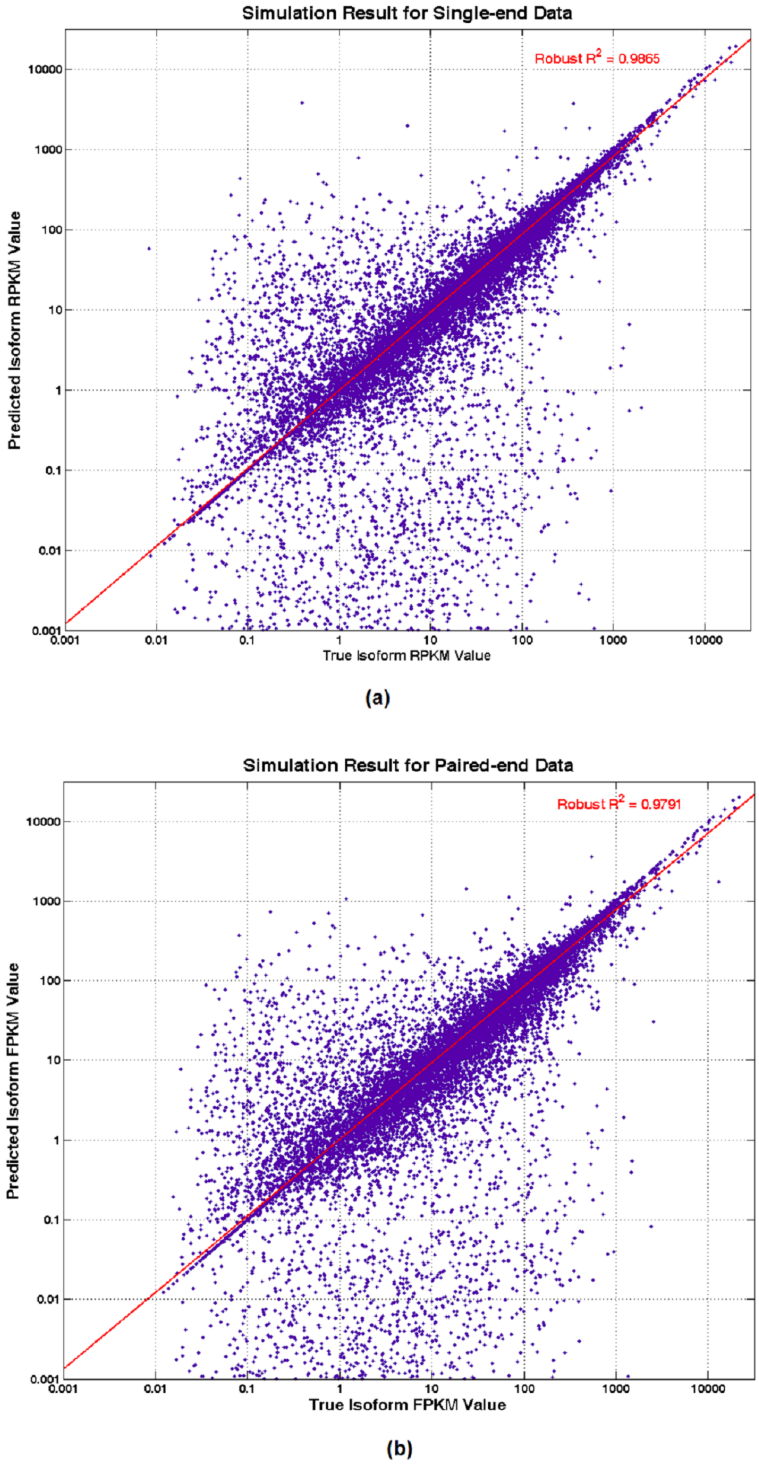


Figure 2.3: Simulation studies of RAEM algorithm [22].

within one gene, the length of the unique exon is shorter than the read length, and so on.

We investigated the transcripts whose abundances were not accurately estimated by RAEM. The transcripts are considered as outliers if that meet $|log10(trueRPKM) - log10(predictedRPKM)| \geq 1$. There are about 10% of transcripts falling into this category. We further examined these outliers. We found that nearly 74% of outlier transcripts belong to genes with more than 5 annotated transcripts and about 41% of outlier transcripts have at least one exon whose length is less than 50 bases. This analysis clearly demonstrates a limitation of RAEM in estimating transcript abundance; however, it is predominantly accurate and effective for around 90% of transcript isoforms in the transcriptome.

2.2 RNA-Seq Analysis of Isoform-level microRNA-155 Target Prediction

2.2.1 Introduction

The regulation of gene expression by microRNAs is a fundamental mechanism for controlling many biological processes. Thus far, more than 1000 microRNA's have been discovered in human cells using either computational or experimental approaches (miRBase [41], release 16, Sept. 2010). The gene encoding the microRNA, microRNA-155, was classified as an oncogene many years before it was identified as a microRNA and is now among the most highly implicated microRNAs in cancer. Despite its link to hematologic and other cancers, there is currently little information regarding direct isoform targets or pathways through which microRNA-155 signals to promote the tumor phenotype.

Over the years, an array of computational approaches have been developed to predict microRNA target sites and these methods have been useful for guiding investigations towards the function of microRNAs [90]. These approaches are roughly divided into rule-based and data-driven approaches [132]. Earlier methods are largely rule-based, predicting microRNA targets as a function of simple discriminative rules derived from features of experimentally validated targets. For example, miRanda [30], DIANA-microT [74], TargetScan [62] and PicTar [56] are mainly based on scanning for conserved 7-mer/8-mer seeds combined with free energy calculations of the RNA-RNA duplex. Latter methods were developed which are more data-driven, such as miTarget [54] and NBmiRTar [131], where machine learning-based approaches were applied to train a classifier that is able to discriminate true microRNA targets from false targets using sequence features.

An alternative data-driven approach is to use 3'-expression microarrays to quantify transcriptomes. In this approach, microRNA targets are predicted by calling significantly down-regulated genes between microRNA over-expressing cell lines and the respective isogenic wild type cell lines [19, 69, 120]. Gene expression based target prediction approaches, (e.g., GenMiR++ [47]), were found to outperform many rule-based approaches, such as [62]. More importantly, the gene expression based approach allows for the discovery of context specific (cell type specific) microRNA target repertoires and this context specific targetome can be related back to the biological processes implicated by the global analysis of the respective microarray experiments. Despite this advantage over purely computational approaches, the intrinsic limitations of the 3'-expression microarrays (such as non-specific hybridization, signal saturation and excessive noise) significantly compromise the performance of microarray based microRNA target prediction.

The advent of Next-Generation Sequencing (NGS) technologies provides new opportunities to profile transcriptomes and microRNA targetomes at base-wise resolution. In our recent work [129], we sequenced the transcriptome of microRNA-155 expressing cells using an Illumina Genome Analyzer II. Our RNA-seq data contains more than one hundred million single-ended 50-mer short reads generated from both wild type Mutu I cells (control) and Mutu I cells expressing microRNA-155 (case). We then developed a computational pipeline to analyze microRNA-155 transcriptome and targetome regulation by performing gene-level down-regulation analysis combined with 7-mer/8-mer seed evidence in 3'-UTR regions. Our analysis yielded a much larger targetome than was previously described using microarray experiments; many predicted microRNA-155 targets were verified by in vitro 3'-UTR reporter assays. Although this analysis was among the first to use RNA-Seq data for microRNA target prediction, this approach did not sufficiently exploit the full value that RNA-Seq data has to offer - that is, using gene structure information derived from the RNA-Seq data to assess isoform specific microRNA regulation. Based on the isoform-level analysis described here, we propose that microRNA targets are more appropriately predicted and characterized at the isoform-level.

On a more general level, we believe that the term, "isoform" or "transcript" may be a more appropriate concept than "gene" in transcriptome studies since the isoform is the ultimate effecter of microRNA responses (as well as many other biological processes). Further, recent studies have shown that microRNA targeting is not limited to the 3'-UTR [61], further emphasizing the need for microRNA target prediction based on the isoform-level.

Genome-wide analysis of transcriptomes and targetomes at the isoform-level is needed, not only for microRNA target prediction but also for many other genomics research areas, such as biomarker discovery, cancer classification, biological pathway

analysis and network reconstruction. The problem itself can be quite challenging since the base-wise gene expression signal from RNA-seq data is often accumulated from a mixture of coexisting isoforms in the living cell. The development of computational algorithms to deconvolve the gene expression signal emitted from each splicing isoform is not a trivial task.

A number of computational approaches have recently been developed to characterize and quantify transcriptomes at the isoform-level (e.g. [9, 43, 64, 50, 92, 114]). These approaches quantify isoform levels of transcripts either annotated in the alternative splicing databases such as those from the UCSC (University of California, Santa Cruz) and Alternative Splicing and Transcript Discovery (ASTD) resources or predicted by short read assembly. However, Our computational approach is among the first batch to predict microRNA targets at the isoform-level. Figure 2.4 shows the workflow of the microRNA target prediction analysis pipeline. We describe each step in details in the following sections.

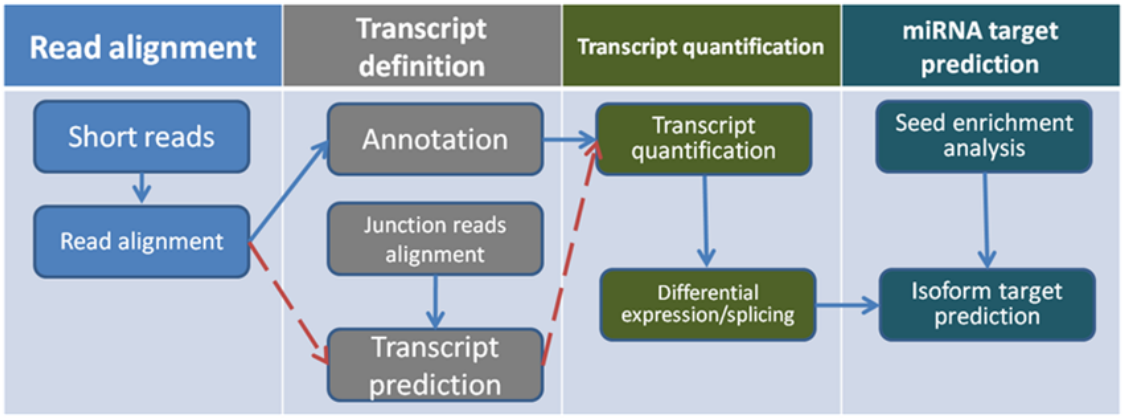


Figure 2.4: The workflow of the microRNA target prediction analysis pipeline [22].

2.2.2 RNA-Seq Data set and Preprocessing

The Burkitt's lymphoma cell line, Mutu I, was retrovirally transduced in duplicate with either a control or a microRNA-155 expressing retrovirus. microRNA-155 real time RT-PCR analysis showed at least 100,000 fold higher expression in microRNA-155 transduced pools relative to control transduced pools [129]. Despite these elevated levels, microRNA-155 expression in transduced Mutu I cells was slightly less than that observed in several activated B-cell lines that naturally express microRNA-155 [129]; arguing against supra-physiological expression of microRNA-155 in transduced Mutu I cells. The transcriptomes of the wild-type Mutu I cell line (6 replicates) and the microRNA-155 expressing Mutu I cell line (6 replicates) were deep sequenced using Illumina Genome Analyzer II with a read length of 50 (NCBI Short Read Archive, Accession Number SRA011001)[129]. For each biological or technical replicate, around 10 million single-ended short reads were generated. Short reads were initially aligned to the reference genome (hg19/GRCh37) using Novoalign (http://www.novocraft.com). We used standard parameter settings to build an index (novoindex) and to run Novoalign. The alignment results were saved in the SAM format and parsed using SAMMate (http://sammate.sourceforge.net/) [128] to calculate gene-level abundance.

2.2.3 Detection of Significantly Down-regulated Gene and Transcript

For isoform-level, we applied RAEM on the RNA-Seq data sets to estimate relative abundance of each transcript. Since the gene-level and isoform-level abundance results we obtained are in the format of continuous numbers, we used a shrinkage t-test

[115] to perform differential expression analysis and call significantly down-regulated genes or transcripts.

2.2.4 Genome-wide Seed Enrichment Analysis

Although there are exceptions, microRNA targeting is primarily guided by 7-mer or 8-mer seeds in a gene region, (usually within the 3'-UTR but sometimes within the 5'-UTR, or an exon). For our analysis, we consider seed enrichment as a necessary condition for target prediction. A single 7-mer or 8-mer seed in a long genome region tends to be less likely to be a microRNA target than many seeds in a short genome region.

We used Pearson's chi-square test to quantify the seed enrichment for each genome region. Basically, we calculate a chi-square test statistic, which quantifies how much the observed seed counts deviate from the expected seed counts in a given genome region. Larger values of the chi-square test statistic (small p-values) will reject the null hypothesis of non-enrichment. The raw p-values of the chi-square test will then be adjusted using the stringent Bonferroni's procedure.

2.2.5 Results: Validation Studies and A Case Study

Our validation studies were carried out using in-house results from 149 different 3'-UTR reporter plasmids containing a spectrum of microRNA-155 seed types, configurations, and potency [120]. The rationale of selecting these 3'-UTRs for in vitro assays is based on current microRNA target database. The 149 genes analyzed in the current study at the 3'-UTR reporter level were selected from a wider panel of 170 such 3'-UTR reporters based on adherence to the following three criteria: (1) the expression estimated from RNA-seq experiments is above 0.5 RPKM at the

gene-level. (2) the expression estimated from our isoform-level approach, RAEM, is above 0.2 RPKM. (3). the genes exhibit a 7-mer/8-mer seed enrichment (adjusted p-value ≤ 0.05) in their 3'-UTR region at the isoform-level. Using the corresponding 3'-UTR reporter data from this set of genes, we tested our isoform-level approach and compared the results to those obtained using the gene-level approach [129]. We used relative expression cut-offs 0.8 (relative expression meaning expression in Mutu-microRNA-155 vs. Mutu-control cells) to discriminate true targets from false targets. We also used a statistical criterion, i.e. q-value [127], as an auxiliary evaluation parameter.

We compared the microRNA-155 targets predicted by gene-level, isoform-level approaches and 3'-UTR assay. In Figure 2.5, the set of 149 targets were divided into eight distinct categories. Because a full list of true microRNA-155 targets is not available as a gold standard, the eight categories essentially represent all possible outcomes of comparing three approaches to microRNA target prediction, i.e. gene-level, isoform-level and 3'-UTR assay.

Targets predicted by both the isoform-level approach and the 3'-UTR reporter assays but not by the gene-level approach (19 predicted targets). This category best highlights the importance of performing isoform-based assessment of microRNA targeting. Here, the differential expression ratio of the target isoform calculated from the isoform-level analysis is more consistent with the 3'-UTR reporter assay results than it is with the results from the gene-level analysis. A good illustration of where this could have important biological significance is the case of TAF5L. TAF5L has three expressed isoforms (ENST00000366676, ENST00000366675 and ENST00000258281). The 3'-UTR of the isoform ENST00000366675 (abundance proportion 11% - 20%, non-dominant isoform) was tested in our 3'-UTR reporter assay, and was predicted by both the isoform-level approach and the 3'-UTR reporter assay as a microRNA-

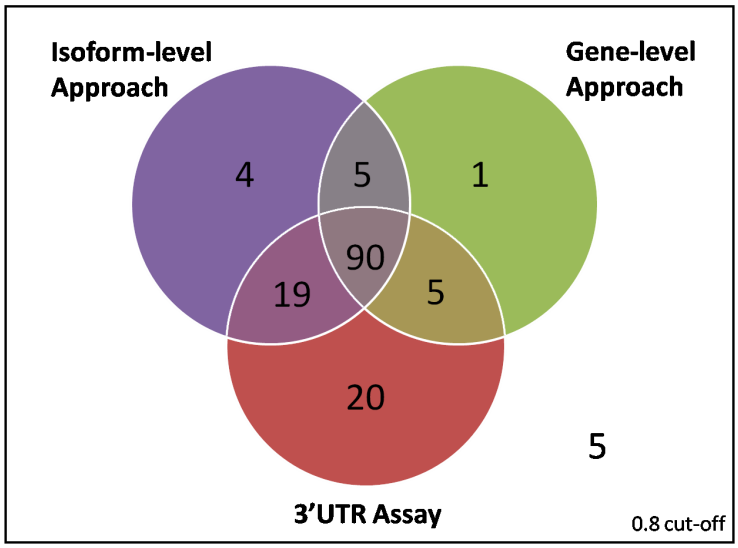
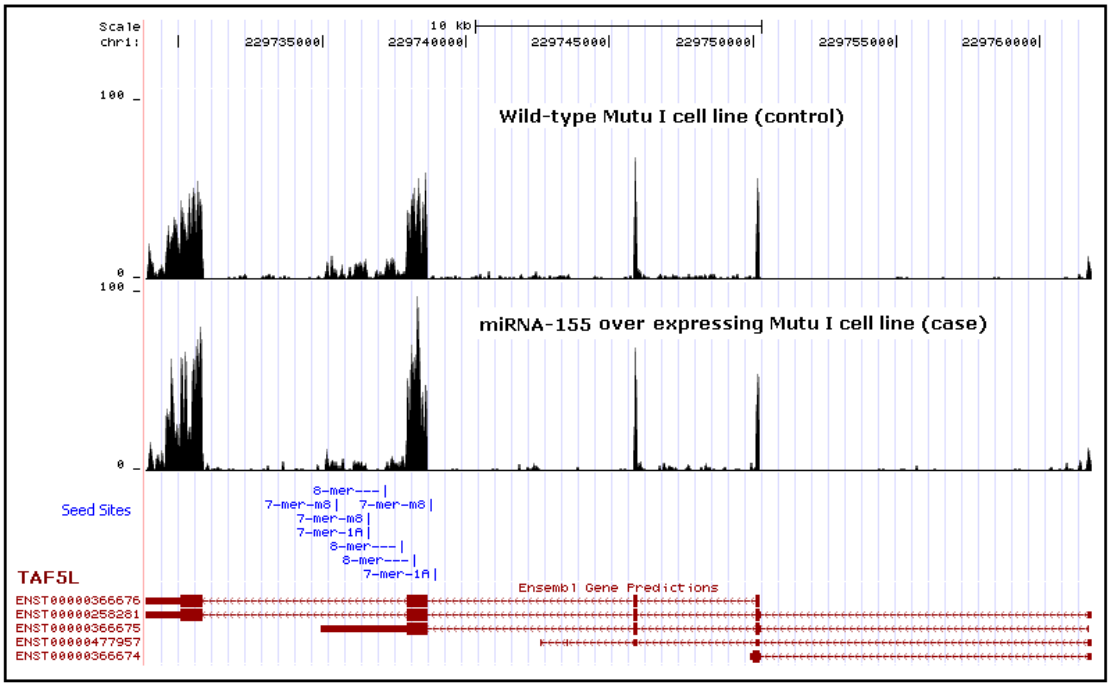


Figure 2.5: The Venn Diagram of the microRNA targets predicted by the three approaches at 0.8 cutoff level of relative expression [22].

155 target (in Figure 2.6). It was not detected by the gene-level approach because this isoform accounts for only 20% of the total gene-level expression in control cells.

To validate that the predominant, unregulated isoform (ENST00000366676) is not responsive to microRNA-155 (as negative control of no repression), we cloned the 3'-UTR of this isoform into a reporter vector and tested it for responsiveness to microRNA-155. As shown in Figure 2.7, while ENST00000366675 again showed inhibition by microRNA-155, the ENST00000366676 isoform was not responsive. To further validate the isoform specific differences in expression at the endogenous RNA level, real time RT-PCR analysis was carried out on microRNA-155 expressing versus control cells using isoform specific PCR primers. As shown in Figure 2.7, RT-PCR demonstrated concordance with the isoform-level analysis of the RNA-seq data. Since the amino acid composition of the proteins expressed from these two isoforms is different at the carboxyl terminus, the isoform specific regulation of one of these isoforms can have a significant regulatory impact on the TAF5L interactome and consequently, TAF5L function.



	Isoform Proportion & Gene Level Expression				Isoform Relative Expression	Seed Enrichment Anysite	Gene Relative Expression	3'UTR Assay
TAF5L	Cntl	Gene RPKM in Cntl	miR155	Gene RPKM in miR155	miR155/Cntl	Location & p-value		
ENST00000366676	76%	20.77	58%	21.02	0.79 (q-value = 9.29%)	Coding, 7.51E-44	1.01 (q-value = 7%)	0.30
ENST00000366675	20%		11%		0.58 (q-value = 0%)	3'UTR, 0		
ENST00000258281	4%		30%		7.27 (q-value = 14.5%)	Coding, 7.51E-44		
ENST00000477957	0%		0%		/	No Enrichment		
ENST00000366674	0%		0%		/	No Enrichment		

Figure 2.6: An example of isoform target predicted jointly by the isoform-level approach and the 3'-UTR assay (Gene TAF5L) [22].

2.2.6 Conclusion

Due to its importance, computational prediction of microRNA targets has been well-studied. However, the existing rule-based, data-driven and expression profiling approaches to target prediction are mostly approached from the gene-level. Gene is a unit of heredity in a living cell that is used extensively in genetics but is becoming a less appropriate concept in transcriptome and targetome research. Here we propose the use of splicing isoform as a more appropriate concept for microRNA target

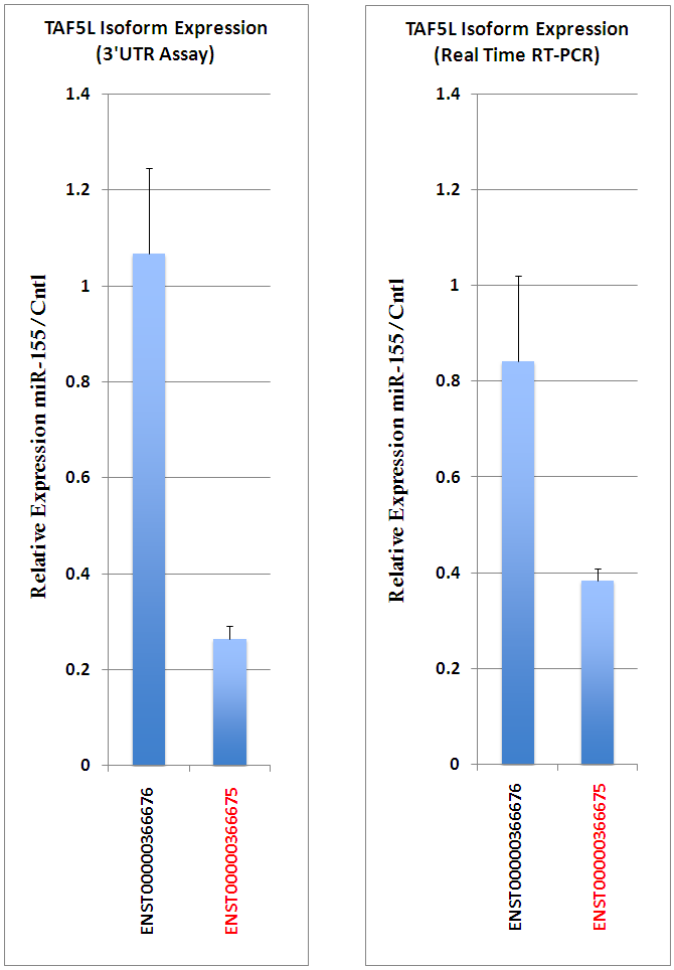


Figure 2.7: Quantitative RT-PCR and 3'UTR reporter assay of the TAF5L isoform relative expression [22].

prediction and other genomics research, since it is the isoform that is the ultimate effector of biological outcomes.

Before the emergence of the deep sequencing technology, exon and tiling microarrays allowed for the analysis of transcriptomes at the isoform-level. The widespread use of these two microarray platforms were limited, however, by intrinsic technological limitations such as resolution, coverage, and signal saturation etc. The advent of deep sequencing technology provides, for the first time, an opportunity to profile transcriptomes at base-wise resolution, making it possible to develop compu-

tational approaches to predict microRNA targets at the isoform-level. We believe this work to be one-of-its-kind, as it allows for the prediction of isoform targets that have not been possible with the gene-level approach that we developed previously [129]. Our computational work has provided deeper biological insights into the microRNA targeting mechanisms as evidenced by in vitro 3'-UTR assay validation.

Chapter 3: dSpliceType : A Novel Algorithm and Tool to Detect Various Types of Differential Splicing Events using RNA-Seq [2]

As mentioned in the previous chapters, alternative splicing plays a key role in regulating process during gene expression in higher eukaryotes [52]. More than 90% of human genes are alternative spliced using different types of splicing mechanisms [118], including skipped exon (SE), retained intron (RI), alternative 3' or 5' splice sites (A3SS or A5SS), and mutually exclusive exons (MXE) [52]. With various types of alternative splicing, isoform transcripts concatenating different exons are transcribed and spliced from a single gene. They are then further translated to produce functionally diverse proteins. Studies have shown that dysregulation of alternative splicing events may lead to various human diseases [52, 75, 119], transcriptome changes between healthy and diseased cells and different stages in cellular development and differentiation [17, 87, 117, 118]. Therefore, efficient and effective algorithms and computational tools for detecting differentially spliced genes and more importantly various types of differential splicing events associated with diseased-specific conditions are urgently needed. New biological insight may be generated to understand the pathological consequences of diseased cell development and differentiation and to eventually identify potential biomarkers for human life-threatening diseases [17, 118, 119].

Over the last decade, expressed sequence tags (EST) and DNA microarrays were widely used to detect differential splicing between transcriptomes by comparing

[2]The content in this chapter is largely derived from original author text and contributions found in [24, 25].

the expression abundances on known gene exons or exon-exon junctions [34, 57, 101, 125, 126, 130]. However, due to limitations of the technologies, the accuracy and precision of detecting differential splicing events were not highly satisfactory [34, 57, 130]. Recently, high-throughput RNA-Seq technologies show promise to interrogate higher eukaryotic transcriptomes more accurately [121]. With millions of short reads directly sequenced from mRNAs at nucleotide base-pair resolution, RNA-Seq not only can be used to estimate relative abundances at both gene and transcript levels [3, 94, 113], but it is also powerful to accurately detect differential splicing [113].

3.1 Novelty of dSpliceType

In general, differential splicing refers to the difference in the relative abundance or proportion of the individual transcripts in a gene between different conditions [46]. Currently, several computational methods exist to detect differentially spliced genes using RNA-Seq. Intuitively, a natural idea is to estimate relative transcript proportions/abundances followed by a statistical test within a gene between conditions to quantify the differences. This type of methods, such as Cufflinks/Cuffdiff [113], [37] and the method [23], is powerful. However, they rely on accurate estimation of transcript abundances, which is a non-trivial problem because of positional and sequence-specific biases from RNA-Seq platforms and read uncertainty. Other methods detect differentially spliced genes by comparing read counts either on all exons within a gene, such as SplicingCompass [5], FDM [102] and MMD [107], or on a single exon, e.g. DEXSeq [4]. These methods can potentially detect differentially spliced genes, but can not specify the regions or associated types of differential splicing events.

Another type of methods is event-based, directly detecting differential splicing events. Certain methods, such as MISO [51] and SpliceTrap [123], focus on the

detection of SE events. More recently, MATS [100] estimates exon inclusion level using junction reads and calculates posterior probability for splicing difference using Markov Chain Monte Carlo (MCMC) method. DiffSplice [46] constructs a splicing event using a splice graph, in which exons and junctions are represented by nodes and edges, and a permutation test is applied to detect the significant splicing events. These two methods are capable of detecting multiple types of events, but may not be optimal for events with partially overlapped exons, such as A3SS, A5SS and RI events. Moreover, either MCMC method (MATS) or permutation test (DiffSplice) makes the detection of differential splicing events time consuming.

When a differential splicing event happens, the discrepancy of read coverage signals in the spliced exonic region among samples across two conditions can be easily observed [22, 23]. Therefore, we develop and present a novel and efficient algorithm and tool, dSpliceType, to detect various types of differential splicing events using RNA-Seq. Compared with the existing methods, dSpliceType has the following novel features. First and most importantly, instead of using read counts, dSpliceType is among the first to detect five types of differential splicing events using read coverage signals from both exonic and junction reads. It utilizes sequential dependency of base-wise signals and detect differential splicing events either between two individual samples using an univariate conditional normal model or among multiple replicates using a multivariate conditional normal model. Second, since we observed that sequencing and alignment biases are likely to affect read coverage signals the same way at each exonic nucleotide in both conditions, dSpliceType is expected to significantly reduce biases by taking ratio of normalized RNA-Seq splicing indexes at each nucleotide between two conditions. Third, according to the results of simulation studies and real-world RNA-Seq data analysis, dSpliceType is demonstrated to be an efficient and accurate computational tool to detect various types of differential splicing events

from a large dynamic range of expressed genes, including relatively low abundance genes.

3.2 A Univariate Algorithm for Detecting Differential Splicing Events without Replicate

3.2.1 Overview

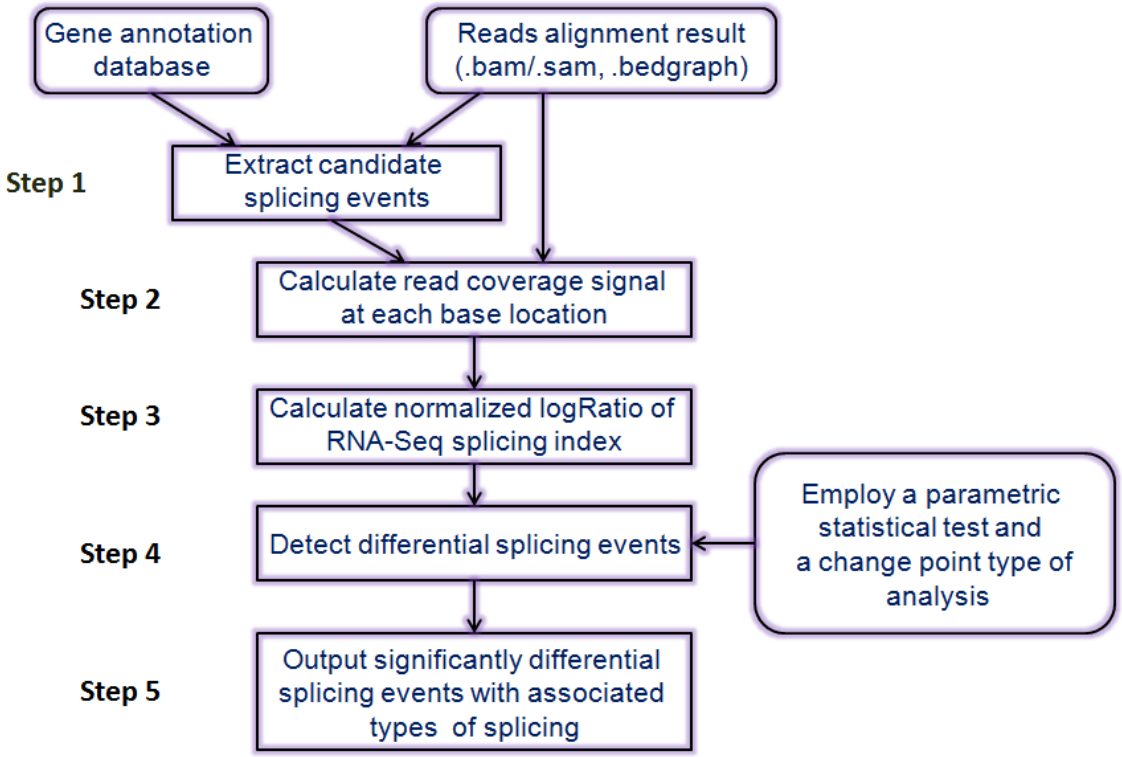


Figure 3.1: The workflow of detecting various types of differential splicing events [24].

dSpliceType is a parametric statistical framework for the detection of differential splicing events using RNA-Seq data. The pipeline of dSpliceType is demonstrated in Figure 3.1. In step 1, annotated splicing events are first extracted from a gene annotation database. We then check the junction reads spanning any two exons associated

with each of annotated splicing events, and keep those supported by junction reads as candidate splicing events. In step 2, at every nucleotide of each candidate splicing event, the read coverage signal is calculated from read alignment result stored in .bam/.sam and/or .bedgraph format for two samples. In step 3, based on the read coverage signals, the normalized logRatio of RNA-Seq splicing indexes at each nucleotide is calculated between two samples. In step 4, a parametric statistical hypothesis test on a conditional normal distribution, capturing sequential dependency of base-wise read coverage signals, is employed, followed by a change-point type of analysis using the Schwarz information criterion on the normalized logRatio of RNA-Seq splicing indexes on each candidate splicing event to detect differential splicing event. In step 5, a p-value is calculated for each candidate splicing event, and the method adjusts the raw p-value using the stringent Bonferroni's procedure. In the following sections, we describe the models for detecting differential splicing events from both two individual samples and multiple replicates between two conditions in the following sections.

3.2.2 Extracting Candidate Splicing Events

Figure 3.2 demonstrates the strategies that our method extracts splicing events for the five well-known types of alternative splicing from gene annotation database.

For skipped exon (SE) events, we explore every coding exon of a transcript, and examine the relationship of two neighbor exons in other transcripts within the same gene. As shown in Figure 3.2A, if the two neighbor exons (E1 and E3) of the specific exon (E2) in tran A are consecutive in tran B, E1, E2 and E3 are extracted and combined as a SE annotated splicing event with recording the starting and ending locations of three exons.

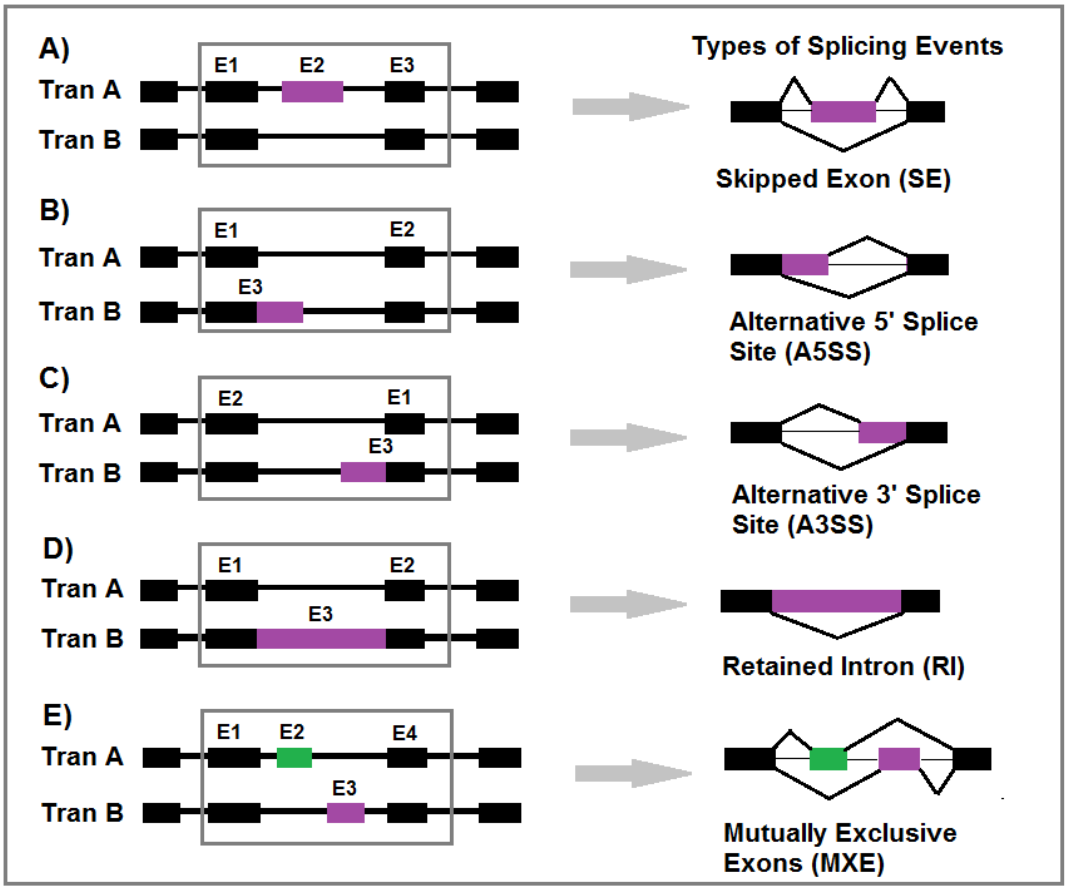


Figure 3.2: The illustration of extracting different types of annotated splicing events [24].

For alternative 5' splice site (A5SS) events, we compare the closest exons towards 5' of a coding exon in two transcripts of a gene. As shown in Figure 3.2B, if the two closest exons (E1 and E3) of a specific exon (E2) have the same starting location in both tran A and tran B, and one is shorter than the other, E1, E2 and E3 are extracted and combined as an A5SS annotated splicing event. Similarly, for alternative 3' splice site (A3SS) events, we compare the closest exons towards 3' of a coding exon in two transcripts of a gene. If the two closest exons (E1 and E3) of a specific exon (E2) have the same ending location and one is shorter than the

other, E1, E2 and E3 are extracted and combined as an A3SS annotated splicing event (Figure 3.2C).

For retained intron (RI) events, we examine two consecutive exons (E1 and E2) in a transcript. If in another transcript within the same gene, there exists an exon (E3) starting at the same location as E1 and ending at the same location as E2, then E1, E2 and E3 are extracted and combined as a RI annotated splicing event (Figure 3.2D).

For mutually exclusive exons (MXE) events, we explore every coding exon of a transcript, and examine the relationship of two neighbor exons in other transcripts within the same gene. If the two neighbor exons (E1 and E4) of the specific exon (E2) in tran A are not consecutive in tran B but with another exon (E3) in between, E1, E2, E3 and E4 are extracted and combined as a MXE annotated splicing event (Figure 3.2E).

To account for the fact that all exons within an annotated splicing event participate in a splicing event, for all five types of annotated splicing events, we keep those supported by junction reads spanning any two exons as candidate splicing events.

3.2.3 Calculating Normalized logRatio of RNA-Seq Splicing Indexes

After extracting different types of candidate splicing events with removed introns, we first calculate the read coverage signal at each nucleotide among three or four exons depending upon the candidate splicing event for two samples, respectively. Second, to compare exonic read coverage signals in terms of alternative splicing, we calculate, for each candidate splicing event of each sample, the RNA-Seq splicing

index at each nucleotide, denoted as SI_i, as given in equation (1). A similar splicing index has been used in microarray studies [126]. We normalize the read coverage signal at the ith nucleotide c_i by the summation of base-wise read coverage signals divided by the total length of the two shared exons of the candidate splicing event (exons in black color as shown in Figure 3.2) as

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

in which le_l and le_r are the length of the left and the right exons, respectively. Finally, we calculate logRatio of the normalized RNA-Seq splicing index SI_i along each candidate splicing event between two samples, which is denoted as $\log\left(SI_{\text{sample1}_i}/SI_{\text{sample2}_i}\right)$.

The advantage of taking ratio of normalized RNA-Seq splicing indexes at each nucleotide location of two samples is to reduce the effect of sequencing and alignment biases from RNA-Seq technology. It is based on the assumption that these biases are more likely to affect read coverage signals at same nucleotide locations on both samples in the same way.

3.2.4 Detecting Differential Splicing Events

The univariate conditional normal distribution model for the normalized logRatio of RNA-Seq splicing index

The normalized logRatio of RNA-Seq splicing index $\log\left(SI_{\text{sample1}_i}/SI_{\text{sample2}_i}\right)$ around zero indicates no read coverage change at the locus, while $\log\left(SI_{\text{sample1}_i}/SI_{\text{sample2}_i}\right)$ smaller or greater than zero indicates read coverage change between the two samples at the locus.

We denote the normalized $\log\left(SI_{\text{sample1}_i}/SI_{\text{sample2}_i}\right)$ at the ith nucleotide along the candidate splicing event as X_i. Due to the sequential dependency among the base-wise read coverage signals, X_i depends on m preceding nucleotides. Therefore, for computational simplicity, we incorporate the first order auto-correlation to capture the sequential dependency between X_i and X_{i-1}, which satisfies the following conditions:

$$\mathrm{E}[X_i] = \mathrm{E}[X_{i-1}] = \mu,$$

$$\mathrm{Var}[X_i] = \mathrm{Var}[X_{i-1}] = \sigma^2 \text{ and}$$

$$\mathrm{Corr}[X_i, X_{i-1}] = \rho \text{ or } \mathrm{Corr}[X_i - \mu, X_{i-1} - \mu] = \rho,$$

and we assume that [49]

$$X_i | X_{i-1} = x_{i-1} \sim N(\mu + \rho(x_{i-1} - \mu), \sigma^2(1 - \rho^2))$$

. Let $\mu' = \mu + \rho(x_{i-1} - \mu)$ and $\sigma'^2 = \sigma^2(1 - \rho^2)$, then $\{X_i | X_{i-1}\}$ can be considered as a series of conditional normal random variables from $N(\mu'_i, \sigma'^2_i)$, for i = 1,...,n, where n is the total exonic length of the candidate splicing event. If no differential splicing happens, μ'_i and σ'^2_i are assumed to be two constant values μ' and σ'^2; while deviations from the constant mean and variance parameters in the spliced region may indicate a differential splicing event.

The hypothesis testing

The identification of differential splicing event 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 variance parameters in the series of $\{X_i|X_{i-1}\}$ [13, 14]:

$$\mathrm{H}_0 : \mu'_1 = \mu'_2 = ... = \mu'_n = \mu' \text{ and}$$

$$\sigma'^2_1 = \sigma'^2_2 = ... = \sigma'^2_n = \sigma'^2 \tag{2}$$

versus the alternative:

$$\mathrm{H}_1 : \mu'_1 = ... = \mu'_{p_1} \neq \mu'_{p_1+1} = ... = \mu'_{p_2} \neq \mu'_{p_2+1}$$

$$= ... = \mu'_{p_q} \neq \mu'_{p_q+1} = ...\mu'_n \text{ and}$$

$$\sigma'^2_1 = ... = \sigma'^2_{p_1} \neq \sigma'^2_{p_1+1} = ... = \sigma'^2_{p_2} \neq \sigma'^2_{p_2+1}$$

$$= ... = \sigma'^2_{p_q} \neq \sigma'^2_{p_q+1} = ...\sigma'^2_n \tag{3}$$

where n is the total exonic length of the candidate splicing event. Under the null hypothesis(2), μ' and σ'^2 are the unknown constant mean and variance; under the alternative hypothesis(3), $1 < p_1 < p_2 < ... < p_q < n$ and $p_1, p_2, ..., p_q$ are the locations of unknown change points.

The null hypothesis(2) relates to no changes in the mean and variance of the conditional normal distribution from the sequence $\{X_i|X_{i-1}\}$, and the alternative hypothesis(3) indicates that multiple changes exist in the parameters of mean and variance. However, based on five different types of candidate splicing events, we consider only two change point locations, which are the ending locations of two exonic regions, for the candidate splicing events of SE, A3SS, A5SS and RI; and three change point locations, which are the ending locations of three exons, for MXE. We then test the null hypothesis(2) versus the new alternative hypothesis(4) or (5) shown as below.

For SE, A3SS, A5SS and RI,

$$\mathrm{H}_1 : \mu'_1 = ... = \mu'_i \neq \mu'_{i+1} = ... = \mu'_j \neq \mu'_{j+1} = ... = \mu'_n \text{ and}$$

$$\sigma'^2_1 = ... = \sigma'^2_i \neq \sigma'^2_{i+1} = ... = \sigma'^2_j \neq \sigma'^2_{j+1} = ... = \sigma'^2_n, \tag{4}$$

where i and j, $1 < i < j < n$, are the two ending locations of the left common exon and the spliced exon/exonic region along the candidate splicing event. For each candidate splicing event of the four types, a significant differential splicing event is detected when the null hypothesis(2) at a given significant level α is rejected, and both i and j are the ending locations of the left common exon (in black) and the spliced exon or spliced exonic region (in purple), respectively, in Figure 3.2.

For MXE,

$$\mathrm{H}_1 : \mu'_1 = ... = \mu'_i \neq \mu'_{i+1} = ... = \mu'_j \neq \mu'_{j+1}$$

$$= ... = \mu'_k \neq \mu'_{k+1} = \mu'_n \text{ and}$$

$$\sigma'^2_1 = ... = \sigma'^2_i \neq \sigma'^2_{i+1} = ... = \sigma'^2_j \neq \sigma'^2_{j+1}$$

$$= ... = \sigma'^2_k \neq \sigma'^2_{k+1} = ... = \sigma'^2_n, \tag{5}$$

where i, j and k, $1 < i < j < k < n$, are the three ending locations of the left common exon and the two spliced exons along the candidate splicing event of MXE. For each MXE candidate splicing event, a significant differential splicing MXE event is detected when the null hypothesis(2) at a given significant level α is rejected, and i, j and k are the ending locations of the left common exon (in black) and the two spliced exons (in purple and green), respectively, in Figure 3.2.

The Schwarz information criterion

In order to test the null hypothesis(2) against the alternative hypothesis(4) or (5), the Schwarz information criterion (SIC)-based method [98] is employed. In general, the SIC is determined by the maximum likelihood function of a model, the number of the estimated parameters as well as the sample size. The model with the minimum SIC indicates the best model for data fitting. Thus, the hypothesis testing can be converted into selecting a model such that the null hypothesis(2) refers to a model without change of mean and variance parameters, while the alternative hypothesis(4) or (5) refers to models with different means and variances specified by two or three change points.

We denote SIC(n) as the SIC corresponding to the null hypothesis(2), which is derived as:

$$\begin{aligned} SIC(n) &= -2\log L_0(\hat{\mu}', \hat{\sigma}'^2, \hat{\rho}) + 3\log n \\ &= n\log 2\pi + n\log \hat{\sigma}'^2 + n + 3\log n \end{aligned}$$

where $\log L_0(\hat{\mu}', \hat{\sigma}'^2, \hat{\rho})$ is the maximum log likelihood function with respect to $H_0(2)$, and $\hat{\mu}'$, $\hat{\sigma}'^2$, and $\hat{\rho}$ are the MLEs of μ', σ'^2, and ρ under H_0, respectively.

Corresponding to $H_1(4)$ 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 $j, 2 \leqslant i, j \leqslant n-2$, is derived as:

$$\begin{aligned} SIC(i,j) &= -2\log L_1(\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3) + 9\log n \\ &= n\log 2\pi + i\log \hat{\sigma}'^2_1 + (j-i)\log \hat{\sigma}'^2_2 \\ &+ (n-j)\log \hat{\sigma}'^2_3 + n + 9\log n, \end{aligned}$$

where $\log L_1(\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3)$ is the maximum log likelihood function with respect to $\mathrm{H}_1(4)$, and $\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3$ are MLEs of corresponding parameters for three models specified by two change points i and j under $\mathrm{H}_1(4)$, respectively.

Similarly, corresponding to $\mathrm{H}_1(5)$ with three change points i, j and k, the SIC for differential splicing events of MXE , denoted by SIC(i, j, k) for fixed i, j and $k, 2 \leqslant i, j, k \leqslant n-2$, is derived as:

$$\begin{aligned} SIC(i,j,k) &= -2\log L_2(\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\mu}'_4, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\sigma}'^2_4, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4) + 12\log n \\ &= n\log 2\pi + i\log\hat{\sigma}'^2_1 + (j-i)\log\hat{\sigma}'^2_2 \\ &\quad + (k-j)\log\hat{\sigma}'^2_3 + (n-k)\log\hat{\sigma}'^2_4 + n + 12\log n, \end{aligned}$$

where $\log L_2(\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\mu}'_4, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\sigma}'^2_4, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4)$ is the maximum log likelihood function with respect to $\mathrm{H}_1(5)$, and $\hat{\mu}'_1, \hat{\mu}'_2, \hat{\mu}'_3, \hat{\mu}'_4, \hat{\sigma}'^2_1, \hat{\sigma}'^2_2, \hat{\sigma}'^2_3, \hat{\sigma}'^2_4, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4$ are MLEs of corresponding parameters for four models specified by three change points i, j and k under $\mathrm{H}_1(5)$, respectively.

Based on the principle of information criterion [14], the null model fits the data better in the sequence of $\{X_i|X_{i-1}\}$ if

$$SIC(n) < SIC(i,j) \quad or \quad SIC(n) < SIC(i,j,k).$$

Otherwise, the model with two change points better fits the data in the sequence of $\{X_i|X_{i-1}\}$ 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 $\{X_i|X_{i-1}\}$ 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.

The test statistic

According to [13], the difference between the SIC scores of the models with and without change points,

$$\Delta_n = SIC(i,j) - SIC(n) \quad or \quad \Delta_n = SIC(i,j,k) - SIC(n)$$

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 (2) against the alternative hypothesis (4) or (5) as

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

where

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

$$b(\log n) = 2\log\log n + \log\log\log n,$$

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

After calculating a p-value for each candidate splicing event, we adjust the raw p-value using the stringent Bonferroni's procedure.

3.3 A Multivariate Algorithm for Detecting Differential Splicing Events with Replicates

3.3.1 Overview

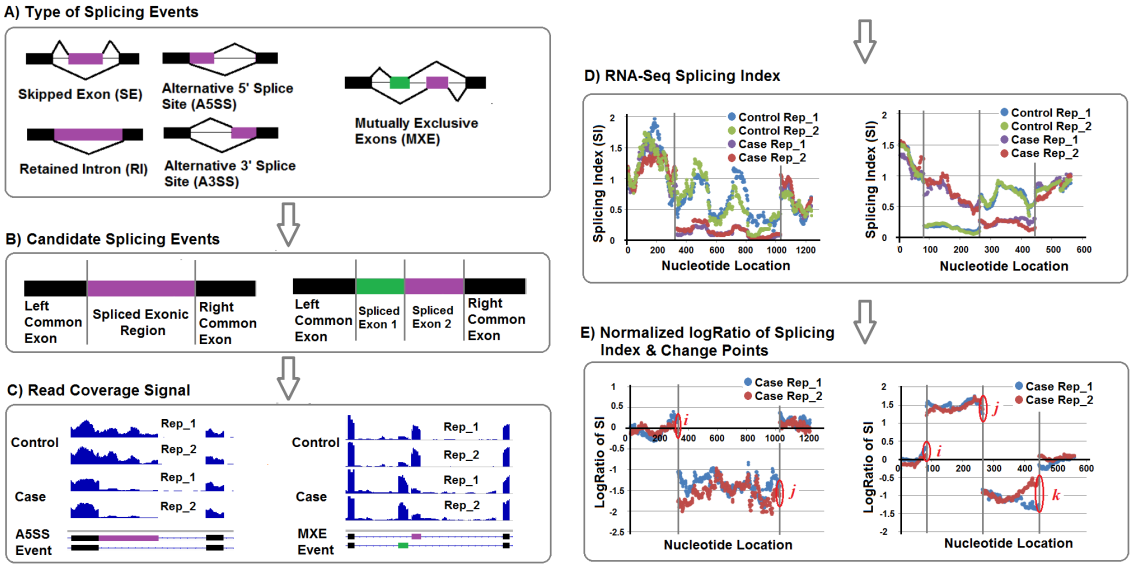


Figure 3.3: The workflow of dSpliceType for detecting various types of differential splicing events with replicates. A) Five most common types of splicing events. Left panel represents SE, RI, A3SS and A5SS events, and right panel represents MXE event. B) Candidate splicing events are compiled by removing 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) and E) 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 [25].

The pipeline is similar to the one demonstrated in Figure 3.1 of Section 3.2. To better illustrate, we present the pipeline in Figure 3.3. First, candidate splicing events are extracted from a gene annotation database supported by junction reads (Figure 3.3A and Figure 3.3B). Second, for each candidate splicing event, read coverage signal

is calculated at each nucleotide location for each replicate in both conditions (Figure 3.3C). Third, based on the read coverage signals, the normalized logRatio of RNA-Seq splicing indexes at each nucleotide location is calculated between two conditions as shown in Figure 3.3D and Figure 3.3E. Finally, for replicates, a series of the normalized logRatio of RNA-Seq splicing indexes along the candidate splicing event is modeled by a multivariate conditional normal distribution, and dSpliceType detects the differential splicing events by employing a change point analysis and a parametric statistical hypothesis test using Schwarz Information Criterion. The raw p-values of multiple tests for differential splicing events are adjusted using the stringent Bonferroni's procedure.

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 3.3A and Figure 3.3B. 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 Section 3.2.2. Novel candidate splicing events can be extracted by incorporating novel junction reads.

Calculating Normalized logRatio of RNA-Seq Splicing Indexes After extracting candidate splicing events, the read coverage signal and the RNA-Seq splicing index at each nucleotide location is calculated in terms of differential splicing for each replicate in both conditions. The RNA-Seq splicing index at the ith nucleotide location is denoted as SI_i given in the following equation. [126] used a similar splicing

index for analysis of differential splicing in microarray studies. 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 3.3B) 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. Finally, the logRatio of normalized RNA-Seq splicing indexes 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.

3.3.2 The Multivariate Conditional Normal Distribution Model for the Normalized logRatio of RNA-Seq Splicing Indexes

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 [27]:

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

where

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

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

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

3.3.3 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 $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ [13]:

$$H_0 : \tilde{\boldsymbol{\mu}}_1 = \tilde{\boldsymbol{\mu}}_2 = ... = \tilde{\boldsymbol{\mu}}_n = \tilde{\boldsymbol{\mu}} \text{ and}$$

$$\tilde{\boldsymbol{\Sigma}}_1 = \tilde{\boldsymbol{\Sigma}}_2 = ... = \tilde{\boldsymbol{\Sigma}}_n = \tilde{\boldsymbol{\Sigma}} \quad (1)$$

versus the alternative:

For SE, A3SS, A5SS and RI,

$$\mathrm{H}_1 : \tilde{\boldsymbol{\mu}}_1 = ... = \tilde{\boldsymbol{\mu}}_i \neq \tilde{\boldsymbol{\mu}}_{i+1} = ... = \tilde{\boldsymbol{\mu}}_j \neq \tilde{\boldsymbol{\mu}}_{j+1} = ... = \tilde{\boldsymbol{\mu}}_n \text{ and}$$

$$\tilde{\boldsymbol{\Sigma}}_1 = ... = \tilde{\boldsymbol{\Sigma}}_i \neq \tilde{\boldsymbol{\Sigma}}_{i+1} = ... = \tilde{\boldsymbol{\Sigma}}_j \neq \tilde{\boldsymbol{\Sigma}}_{j+1} = ... = \tilde{\boldsymbol{\Sigma}}_n, \tag{2}$$

where i and j, $1 < i < j < n$, are the unknown two locations along the candidate splicing event. For each candidate splicing event of the four types, a significant differential splicing event is detected when the null hypothesis(1) at a given significant level α is rejected, and both i and j are within a small offset of the ending locations of the left common exon (in black) and the spliced exon or spliced exonic region (in purple), respectively, in Figure 3.2.

For MXE,

$$\mathrm{H}_1 : \tilde{\boldsymbol{\mu}}_1 = ... = \tilde{\boldsymbol{\mu}}_i \neq \tilde{\boldsymbol{\mu}}_{i+1} = ... = \tilde{\boldsymbol{\mu}}_j \neq \tilde{\boldsymbol{\mu}}_{j+1} = ... = \tilde{\boldsymbol{\mu}}_k \neq \tilde{\boldsymbol{\mu}}_{k+1} = ... = \tilde{\boldsymbol{\mu}}_n \text{ and}$$

$$\tilde{\boldsymbol{\Sigma}}_1 = ... = \tilde{\boldsymbol{\Sigma}}_i \neq \tilde{\boldsymbol{\Sigma}}_{i+1} = ... = \tilde{\boldsymbol{\Sigma}}_j \neq \tilde{\boldsymbol{\Sigma}}_{j+1} = ... = \tilde{\boldsymbol{\Sigma}}_k \neq \tilde{\boldsymbol{\Sigma}}_{k+1} = ... = \tilde{\boldsymbol{\Sigma}}_n, \tag{3}$$

where i, j and k, $1 < i < j < k < n$, are the unknown locations along the candidate splicing event of MXE. For each MXE candidate splicing event, a significant differential splicing MXE event is detected when the null hypothesis(1) at a given significant level α is rejected, and i, j and k are within a small offset of the ending locations of the left common exon (in black) and the two spliced exons (in purple and green), respectively, in Figure 3.2.

3.3.4 The Schwarz Information Criterion

To test the null hypothesis(1) against the alternative hypothesis(2) or (3), the Schwarz information criterion (SIC)-based method [98] is employed. The model with the minimum SIC indicates the best model for data fitting. Thus, the hypothesis testing can be converted into selecting a model such that the null hypothesis(1) refers to a model without change of mean and covariance parameters, while the alternative hypothesis(2) or (3) refers to 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 :

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

where the log likelihood is

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

So, we have

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

Corresponding to $H_1(2)$ 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:

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

Similarly, corresponding to $\mathrm{H}_1(3)$ with three change points i, j and k, the SIC for differential splicing events of MXE , denoted by SIC(i, j, k) for fixed i, j and $k, m \leqslant i, j, k \leqslant n-m$, is derived as:

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

According to the principle of information criterion [13], the null model fits the data in the sequence of $\{\mathbf{X}_i|\mathbf{X}_{i-1}\}$ better if

$$SIC(n) < SIC(i,j) \quad or \quad SIC(n) < SIC(i,j,k).$$

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.

3.3.5 The Test Statistic

According to [13], the difference between the SIC scores of the models with and without change points,

$$\Delta_n = SIC(i,j) - SIC(n) \quad or \quad \Delta_n = SIC(i,j,k) - SIC(n)$$

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.

3.4 Results

3.4.1 Simulation Studies

Simulated data sets

We evaluated the accuracy of dSpliceType and compared the performance with two existing methods, MATS [100] and Cufflinks/Cuffdiff [113], using simulation studies. FluxSimulator [39] was used to generate 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 in each group.

In each group of RNA-Seq data sets, 8,031 SE, 3,711 A3SS, 3,175 A5SS, 1,661 RI and 1,366 MXE events, corresponding to 40,753 event-related transcripts and 10,275 differentially spliced genes, were simulated from the annotated Ensembl database (version 69). For each splicing event, the spliced exon or region is more than 3 nucleotides. Among 10,275 differentially spliced genes, each gene on average has 4 event-related transcripts expressed and contains multiple splicing events. To simulate various splicing ratios in the five types of differential splicing events between two conditions, FluxSimulator was first used to randomly generate copy numbers of event-related transcripts as expression profiles for control condition, and then we generated transcript expression profiles for case condition by re-ordering the copy numbers of event-related transcripts in each gene of control condition. We used different transcript profiles to generate short reads for replicates in two conditions.

We mapped the simulated RNA-Seq data sets uniquely to the human reference genome (hg19/GRCh37) using Tophat2 [53] and Bowtie2 [59]. 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 [89] 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 10 at least in one condition, 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 3.1 and Figure 3.4 are all true positives, and no false positive is detected by all three methods with their parameter settings.

Simulation results of detecting differentially spliced genes

Table 3.1: Comparison of the differentially spliced genes detected by dSpliceType, MATS and Cuffdiff in 4 groups of simulated data sets. For each method, the highest detection rate is in bold face [25].

# of Reads	**# of Spliced Genes[1]**	**Methods**					
		dSpliceType		**MATS**		**Cuffdiff**	
30M		8,054	78%	7,148	70%	2,086	20%
50M	10,275	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 simulated data sets. M stands for million.

We compared the overall performances of the three computational methods on detecting differentially spliced genes in 4 groups of simulated data sets. We collected

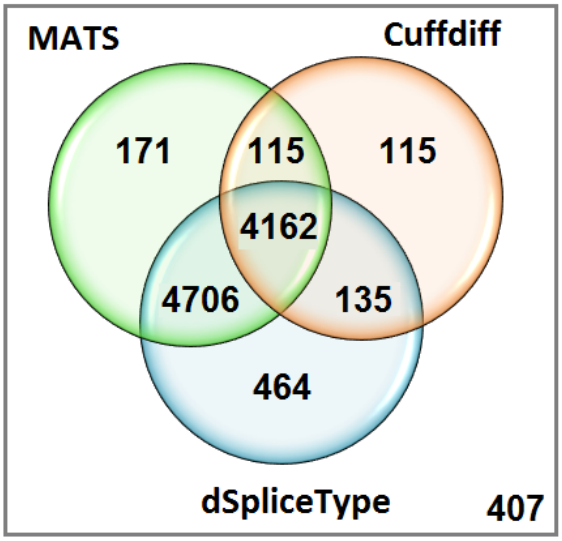


Figure 3.4: The comparison of the detected differentially spliced genes by dSpliceType, MATS and Cuffdiff (200 million simulated data set) [25].

the detected 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 3.1 shows that dSpliceType outperforms the other two methods by achieving the highest numbers and detection rates in all simulated data sets, and dSpliceType can detect 92% of the true differentially spliced genes when the sequencing depth reaches 200 million reads. One possible reason for the lowest detection rate for Cuffdiff is the inaccurate estimation of relative transcript abundances of genes when a gene has many annotated transcripts.

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 3.4, 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 3.2.

Table 3.2: 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) [25].

	# 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.

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 3.4. Therefore, dSpliceType is demonstrated to be able to detect differentially spliced genes in a large dynamic range of expressed genes.

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 gap between them. Table 3.3 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

Table 3.3: 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.

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).

Runtime comparison

Table 3.4: Runtime comparison of dSpliceType, MATS and Cuffdiff in 4 groups of simulated data sets. The shortest runtimes are highlighted in bold [25].

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. M stands for million.

Table 3.4 shows the runtime comparison of the three methods among the 4 groups of simulated 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 [89] and the time of detecting differential splicing events by dSpliceType. The conversion time increases linearly with the number of reads, for example, 36 minutes for 6 samples of 30 million data set and 150 minutes for 200 million data set; while the detection time of dSpliceType does not change much as 25 minutes for 30 million and 32 minutes for 200 million data sets. This is because the total number of nucleotides covered by short reads of 4 groups of simulated data sets are quite similar, and only the values of read coverage signals on those nucleotides are changed when the number of reads increases. Therefore, as shown in Table 3.4, the increase in the number of reads reflects more of the increase in conversion time, not detection time.

3.4.2 Real-world Data Analysis

RNA-Seq data and pre-processing

We applied dSpliceType to a public paired-end Illumina 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 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.

Detection of differential splicing events

dSpliceType detected amount of differential splicing events. We present five differential splicing events detected by dSpliceType with different types of alternative splicing in Figure 3.5, in which MATS can detect four except the A3SS differential splicing event of gene DNAJC10, and Cuffdiff only detected gene CLK4 as a differentially spliced gene.

Each case study of differential splicing events includes the plots of the read coverage signals on the candidate splicing event, the calculated RNA-Seq splicing indexes and the logRatio of RNA-Seq splicing indexes, and the change points i, j or i, j and k at the ending locations of the exons of the differential splicing event. For all the five case studies, it can be seen that the values of logRatio of the RNA-Seq splicing indexes in the shared exonic regions are close to zero, which indicates that no obvious read coverage changes exists in them between the two conditions after

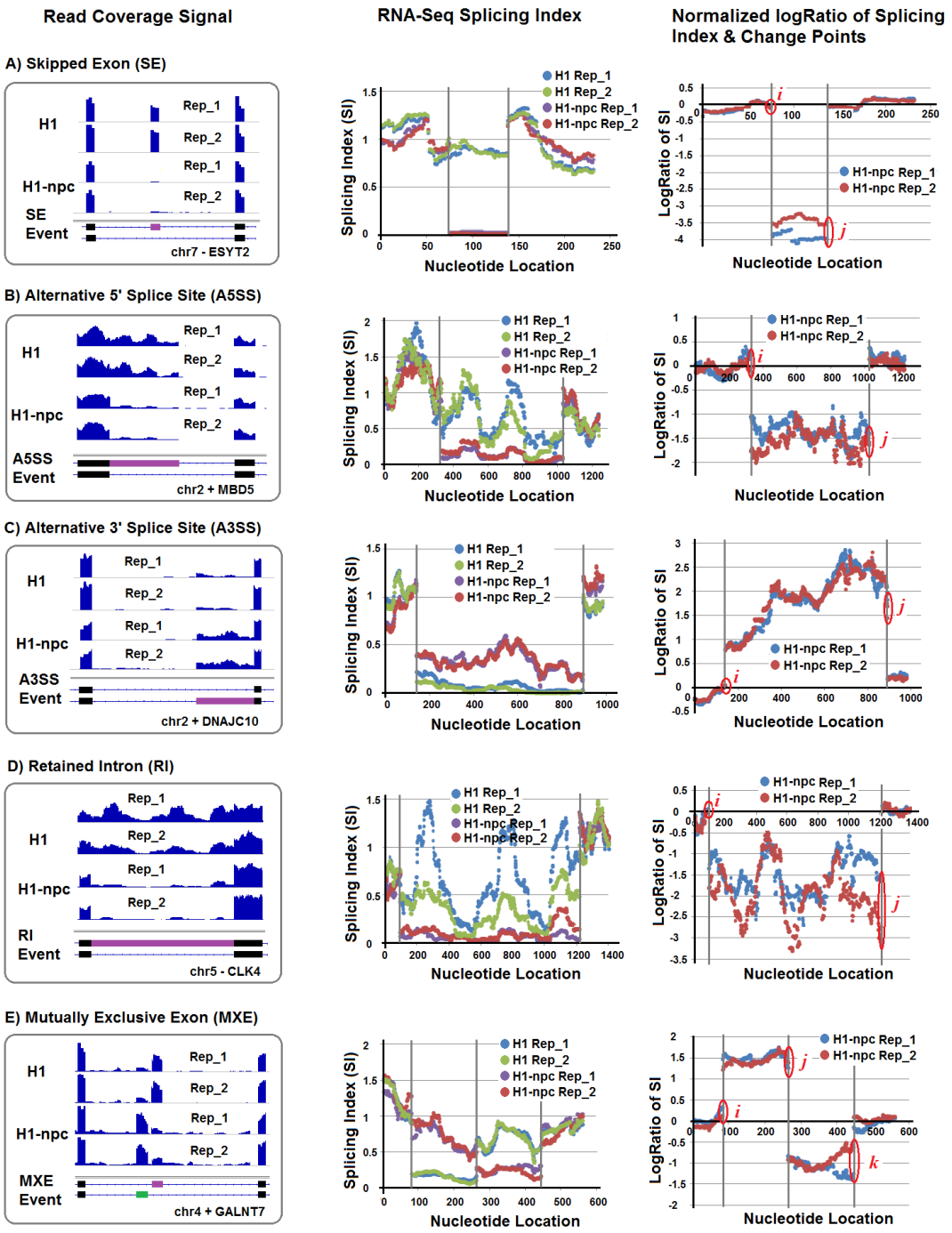


Figure 3.5: The five case studies of detected differential splicing events with replicates by dSpliceType. Each row indicates a case study, and three columns show the plots of read coverage signal, RNA-Seq splicing index, and logRatio of splicing index and detected change points, respectively. (A) A skipped exon (SE) differential splicing event is detected for the gene chr7 - ESYT2. (B) An alternative 5' splice site (A5SS) differential splicing event is detected for the gene chr2 + MBD5. (C) An alternative 3' splice site (A3SS) differential splicing event is detected for the gene chr2 + DNAJC10. (D) A retained intron (RI) differential splicing event is detected for the gene chr5 - CLK4. (E) A mutually exclusive exon (MXE) differential splicing event is detected for the gene chr4 + GALNT7 [25].

normalization. However, the values of logRatio of the RNA-Seq splicing indexes in the spliced exonic regions deviate from zero with different variations regarding to different types of alternative splicing, which reflects the read coverage discrepancy when the differential splicing event happens.

3.5 Discussion and Conclusion

As studies increasingly shift from DNA microarrays, RNA-Seq holds the promise to better interrogate transcriptomes, particularly splicing mechanisms. The method, dSpliceType, is designed specifically to utilize read coverage signals and work with multiple biological replicates.

Compared with read-count based methods, dSpliceType has the following major advantages. dSpliceType detects accurately differential splicing events. Instead of complex model for bias correction, we believe that taking ratio of normalized RNA-Seq splicing indexes between conditions is an efficient way to eliminate the effect of sampling biases from RNA-Seq. We use a model of multivariate conditional normal to capture the sequential dependency of the read coverage signals after normalization and taking logRatio, and detect differential splicing events by comparing SIC scores between models with or without change points.

The read counts based differential splicing methods usually require sequencing at a certain depth. Therefore, the performance of these methods may be limited for genes with low abundances. However, as a read-coverage based method, dSpliceType overcomes this limitation; the detection can be effective as long as the nucleotides of the splicing event are covered by reads, regardless of coverage depth. This is because a sufficient number of per-base coverage signal values in exonic regions can be used to accurately estimate model parameters (i.e., the means and variance covariance

matrices), and the sharp signal changes on exon-exon boundaries of the splicing events can be effectively identified as change points, even if the read coverage is relatively low.

The increasing depth of RNA-Seq allows read-count based methods to detect more differentially spliced genes. However, this is also more computational intensive if the method needs to process every reads and employs an iterative or re-sampling procedure. dSpliceType is not sensitive to the increase in the number of short reads because it provides a closed-form solution using read coverage signals, and the increasing number of reads do not incur extra computational load since they primarily result in elevated values of read coverage signals. Converting read alignment result to read coverage signals is considered to linearly increase the time complexity. Thus, dSpliceType is time efficient.

As a read-coverage based method, dSpliceType can be applied to RNA-Seq data from multiple sequencing platforms, with longer or shorter read lengths, as long as the base-wise read coverage signals are available. dSpliceType is expected to be more powerful with the ever-increasing sequencing coverage depth.

Chapter 4: A RNA-Seq Computational Workflow to Jointly Study Genes with Differential Expression and Differential Splicing [3]

4.1 Introduction of the Computational Workflow

Most of human genes are alternatively spliced. As a result, it drastically increases the diversity of functional proteins. Using RNA-Seq data, there exists a greater potential to better interrogate human transcriptomes. As presented in Chapter 2, we developed an EM-based transcriptome quantification algorithm and tool, RAEM, to estimate relative transcript abundances, which can be used to conduct differential expression analysis at isoform and/or gene levels. Also, in Chapter 3, a novel algorithm and tool, dSpliceType, was developed and presented for detecting differential splicing events using RNA-Seq data. Differential expression or differential splicing analysis may provide novel biological insights; however, one type of analysis only provides difference of human transcriptomes from one biological angle. To better characterize and understand the pathological consequences of serious human diseases, in this chapter, we have developed a computational workflow to jointly study genes from two aspects, differential expression and differential splicing, simultaneously.

Figure 4.1 shows the steps of the computational workflow to jointly study genes with differential expression and differential splicing. The workflow conducts transcript quantification (estimating the relative transcript proportions/abundances) by taking read alignment files (.sam, .bam, or .bedgraph) and gene annotation database (.gtf

[3] The content in this chapter is largely derived from original author text and contributions found in [23].

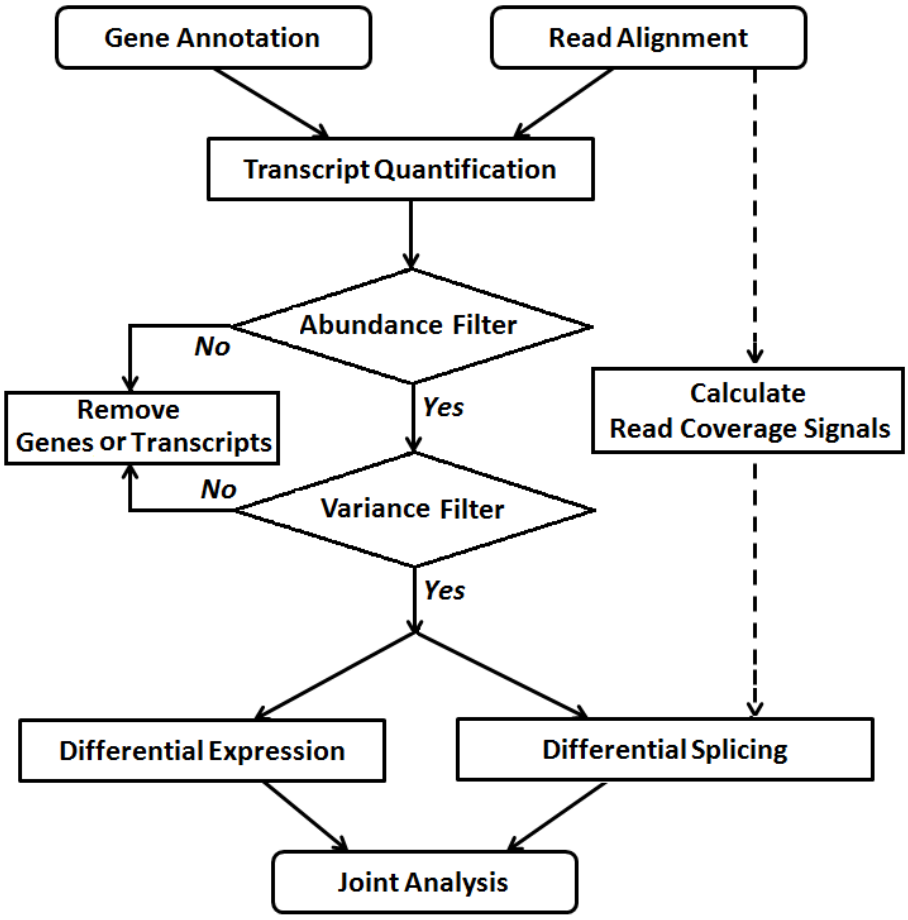


Figure 4.1: The computational workflow of joint study of differential expression and differential splicing.

or .gff file) as input. Gene-level and isoform-level differential expression analysis can be performed based on the estimation of transcript relative abundances using RAEM followed by abundance and/or variance filtering with certain cutoff. Differentially spliced genes can be detected by quantifying the discrepancy of relative transcript proportions using RAEM in each gene across samples using a statistical test or detected directly from read coverage signals using dSpliceType. Finally, we present genes by combining the analysis results of both differential expression and differential splicing. The computational workflow is employed to study a human idiopathic pulmonary fibrosis (IPF) lung disease. We present gene information in two dimensions

in terms of both differential expression and differential splicing, and we are able to detect differential splicing variants from non-differentially expressed genes as potential biomarkers.

4.2 Detecting Splicing Variants from non-Differentially Expressed Genes in a Human Lung Disease

4.2.1 Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive form of chronic lung scarring, which occurs predominantly in older adults and carries a dismal prognosis. Studies indicate that 50% of patients with IPF die within 3 years of diagnosis [72] and that the majority of afflicted patients die from IPF [18]. To date there are no known agents that reduce mortality of IPF and clinical trials are stymied by a dearth of clinically employed biomarkers. Our understanding of the pathogenesis of IPF is far from complete, and to date there has been a lack of powerful, high throughput molecular profiling techniques that permit delineation of the whole transcriptome landscape at high resolution.

Splice variants occur in conjunction with fibrosis in the lung and other organs [29, 36, 79, 99, 110]. Traditional methods use high throughput gene expression profiling techniques, such as microarray, to detect differentially expressed genes at the whole-transcriptome scale. In-depth examination of the splicing of the top ranked genes using lower throughput, but more accurate techniques, such as qRT-PCR [34], can be subsequently performed. These approaches have proven useful, but they do not permit a comprehensive transcriptomic landscape at the level of splicing variants.

The declining cost and increasing throughput of RNA-seq technology provide new opportunities to characterize the highly diverse and complex human transcriptome. Compared with the older tilting and exon arrays, RNA-seq provides abundant signal at base-pair resolution, and promises a better means to identify and quantify splicing variants in the human transcriptome [7, 22, 91, 92, 121]. Examining transcriptomes at the isoform-level allows for detecting differential regulated splicing variants encoded by the non-differentially expressed genes, which may be important but are often hidden from discovery by many older microarray techniques. We apply our method and report the whole transcriptome-scale analysis of differential splicing events in IPF patient samples using RNA-seq. To the best of our knowledge, this is the first study that examines splicing variants from non-differentially expressed genes for human IPF disease.

4.2.2 RNA-Seq Dataset and Preprocessing

Human IPF and control lung specimens were obtained from the NIH Lung Tissue Research Consortium (LTRC). The transcriptomes of 3 IPF patient samples and 3 age-matched controls were deep-sequenced using an Illumina Genome Analyzer II with a read length of 54 bases. This is considered a suitable control group because most patients with IPF have been smokers. For each tissue sample (biological replicate), over 25 million single-end reads were generated and stored in a file with fastq format. The RNA-seq data were submitted to the NCBI Short Read Archive with accession number SRA048904.

We first performed a per base sequence quality check using fastQC Software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). We confirmed the high short read quality of each sample, with the average quality score at each base position

above 35 using fastQC, which is much higher than the recommended threshold of 20. TopHat (v1.0.14) [112] was used thereafter to align short reads that were unique to the human reference genome (release hg19/GRCh37). Default settings and '-g 1' parameter were used. Alignment results were saved in a SAM format. For each sample, between 60% and 70% of reads were uniquely aligned to the reference genome, representing in a sufficient amount of aligned reads for analyses.

4.2.3 Expression Abundance Estimation at both Gene and Isoform Level

Based on the alignment results of all the samples, expression abundance estimation was conducted at both gene and isoform levels using Ensembl ASTD database (version 60). SAMMate [128] (http://sammate.sourceforge.net/), free Graphical User Interface (GUI) software, was employed for gene and isoform quantification. For isoform quantification, we applied the method, RAEM (Reads Assign by Expectation Maximization), reported in [22] and implemented in SAMMate. The output results of SAMMate contain not only expression abundance level measured by RPKM (Reads Per Kilobase of exon or transcript model per Million mapped reads) [78], but also aligned read counts for each gene and isoform, which were used as the input for differential expression analysis using edgeR [94].

4.2.4 Differential Expression Analysis at both Gene and Isoform Level

Firstly, an average RPKM cut off of 1 in both control and case conditions was applied to remove very low-abundance genes. And then, for those isoforms belonging to the remaining genes, we further filtered out very low-abundance isoforms,

which have an average RPKM value less than 1 in both conditions. Finally, edgeR was employed to prioritize the differentially expressed genes and isoforms with FDR values.

4.2.5 Differential Splicing Detection

We applied RAEM to estimate isoform proportions for each gene. With the abovementioned abundance filtering results, in order to identify the most consistent differential splicing events, we also applied the variance filtering at the isoform-level. First, for each isoform of each gene, we multiplied the estimated isoform proportion by 100, e.g. enlarging 90% to 90, and then calculated the enlarged proportion variance for both control and case conditions. If the variance of enlarged proportion of every isoform in both conditions is smaller than 150 (the cutoff of the variance filter), we considered that the gene does not contain too much proportion variance on its isoforms, and keep these genes as candidates. Secondly, for each candidate gene, we constructed an n x 2 matrix. In the matrix, n rows correspond to the isoforms, and two columns correspond to the average of isoform enlarged proportions in control and case conditions respectively. Finally, for candidate genes, to detect differential splicing events between two conditions, we applied the Pearson's Chi-squared test of independence (R function chisq.test) with Yates' correction for continuity, and ranked those genes by FDR values, which are calculated from raw Chi-square p-values using Benjamini-Hochberg procedure [8].

In differential splicing analysis, we primarily focused on the divergence of isoform proportions for each gene across the IPF and control conditions. We examined all the genes as long as their expression abundances are above a certain threshold. In general, the Chi-squared test should be applied on actual count data, e.g. RPKM

value of each isoform. However, directly using read count data has the following limitations: for highly expressed genes (> 100 RPKM), even the minor change of isoform proportions between two conditions can yield significant differential splicing events (significant p-values); on the other hand, for relative low abundance genes (< 10 RPKM), major change of isoform proportions between two conditions can be missed (non-significant p-values). Thus, we used the average of enlarged proportions as pseudo counts to make the p-values comparable across the genes of different abundances and achieve a more robust detection of differential splicing.

4.2.6 Joint Results

Differential Expression and Differential Splicing Analysis

In total, 110,982 protein coding isoforms were annotated, corresponding to 20,560 protein coding genes. After gene and isoform abundance filtering, 13,923 genes and 44,396 isoforms were selected for differential expression analysis at the gene- and isoform-level, respectively.

Although many known genes are differentially expressed, splicing variants can display a characteristic "switch" between major and minor isoforms. In addition to variations in overall gene expression, differential splicing may also be important to fully understand the underlying mechanisms involved in the pathobiology of IPF. Since differential splicing isoforms may play an important role in lung fibrosis [99], we conducted differential splicing analysis at the whole transcriptome scale, investigating those genes in which the proportions of expressed isoforms change (major-minor isoform switch) between control and case conditions.

After abundance and variance filtering, 3,098 genes with more than 1 expressed isoform were left as candidates for differential splicing analysis. Among these, 248

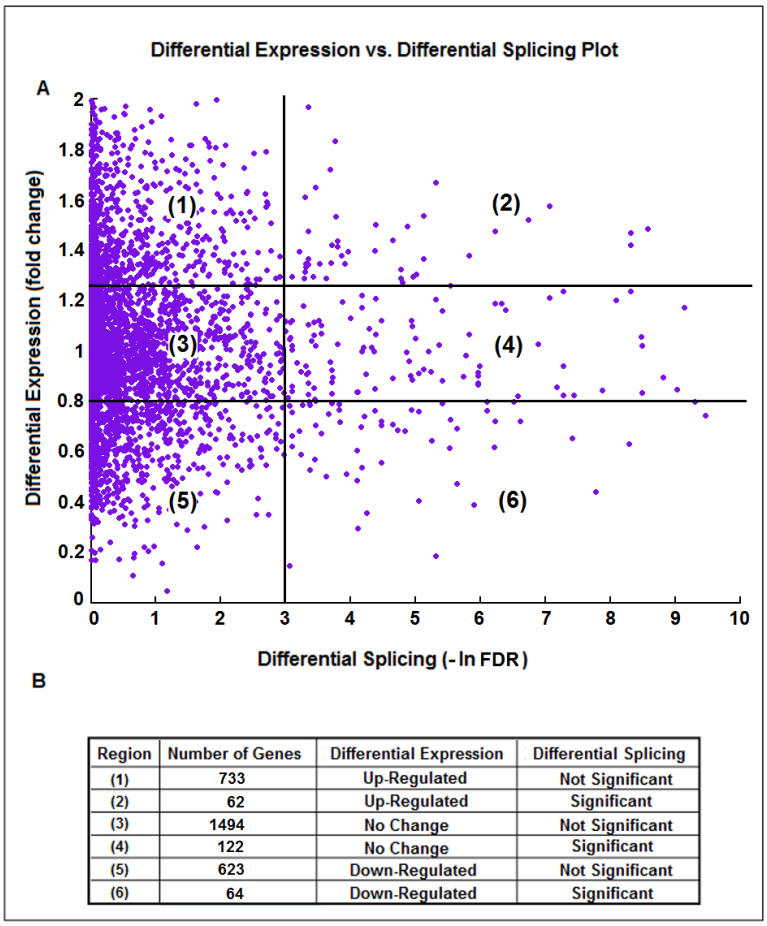
genes have Chi-square test False Discovery Rate (FDR) less than 0.05, and we considered these genes differentially spliced with statistical significance.

Joint Analysis of Differential Expression and Differential Splicing

Although newer microarray technologies, such as exon-junction array and tilting array, enable transcriptomic analysis at the isoform-level [48, 58, 85, 101], the sensitivity and specificity are inherently limited by signal saturation, probe design and non-specific hybridization. Compared with microarray technologies, RNA-seq provides nucleotides sequencing at base-pair resolution, and therefore increases the accuracy of differential expression and differential splicing analyses. Since these two types of analyses examine different aspects of gene expression variation, it is necessary to perform a joint analysis to uncover novel biological events that could not be revealed by each alone. We attempted to identify IPF splicing variants that are consistent among replicates by examining the major-minor switch of isoform proportions within each gene. The combined information for these genes is presented in Figure 4.2.

In Figure 4.2A, 3,098 genes were plotted, with each purple dot corresponding to one gene, after abundance and variance filtering described in the Methods Section. The whole panel is further partitioned into 6 regions and detailed information is shown in Figure 4.2B. For each gene, we define up-regulation as fold change > 1.25, down-regulation as fold change < 0.8 and no change as fold change between 0.8 and 1.25. We also define significant differential splicing as -ln (FDR) > 3 (corresponding to FDR value < 0.05). In Figure 4.2A, most genes fall into region (1), (3) and (5), representing genes without major-minor isoform switches.

The genes located in regions (2), (4) and (6), however, are significantly differentially spliced, and would not be discovered by gene-level analysis. In particular,



Region	Number of Genes	Differential Expression	Differential Splicing
(1)	733	Up-Regulated	Not Significant
(2)	62	Up-Regulated	Significant
(3)	1494	No Change	Not Significant
(4)	122	No Change	Significant
(5)	623	Down-Regulated	Not Significant
(6)	64	Down-Regulated	Significant

Figure 4.2: Joint Study of differential expression and differential splicing between IPF lungs and controls [23].

genes in region (4) are not differentially expressed at the gene level but display significantly differential splicing, so genes in this region represent a novel and previously uncharacteristic region of regulation that warrants further investigation.

A Case Study of Predicted Differentially Spliced Gene

We discuss in detail an example (gene TOM1L1) from region (4) (Figure 4.2A) with strong differential splicing evidence based on their read coverage signal maps.

The case correspond to one type of alternative splicing in IPF without significant expression abundance changes at the gene-level.

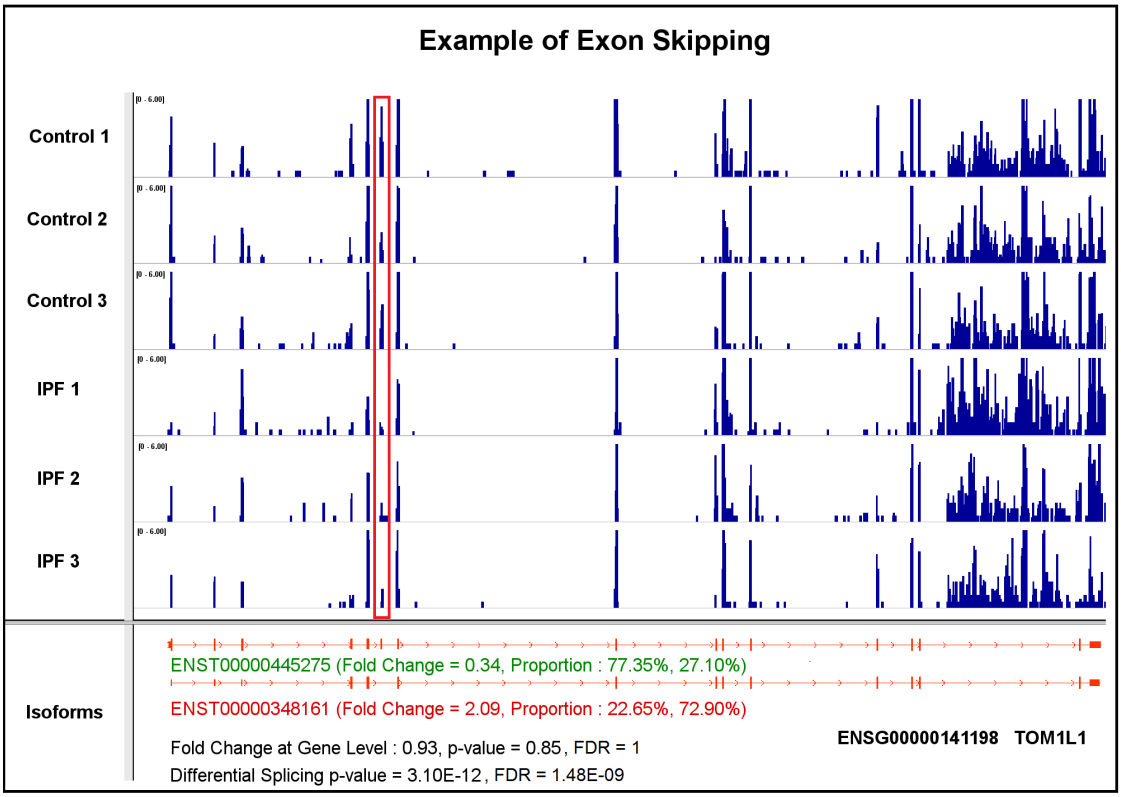


Figure 4.3: A case study of gene TOM1L1 illustrating the skipped exon splicing event using the annotated transcripts [23].

Exon skipping or cassette exon is the most common type of alternative splicing event in eukaryotic species [96]. A representative of this type of alternative splicing is TOM1L1, which has two annotated isoforms: ENST00000445275 and ENST00000348161 (Figure 4.3). The major difference between these two isoforms is that the 6th exon of ENST00000445275 is skipped in ENST00000348161. Importantly, with a 0.93 fold change, the gene is considered to show no differential expression. However, based on our differential expression analysis at the isoform-level, ENST00000445275 is down-regulated and ENST00000348161 is up-regulated, with

0.34 and 2.09 fold changes, respectively. The observed gene-level differential expression ratio (DER) (0.93) represents the mixture of isoform-level DER (0.34 and 2.09). Meanwhile, the isoform proportion of ENST00000445275 decreases from 77.35% in control to 27.10% in IPF cases, while the isoform proportion of ENST00000348161 increases from 22.65% to 72.90%. These differences between control and case condition indicate a high degree of major-minor isoform switches, as the differential splicing FDR value is 1.48E-09. It also reveals the advantage of isoform-level differential expression and differential splicing analysis. The red box highlights the decreased read coverage at the skipped exon from control to case condition as the evidence of the exon skipping.

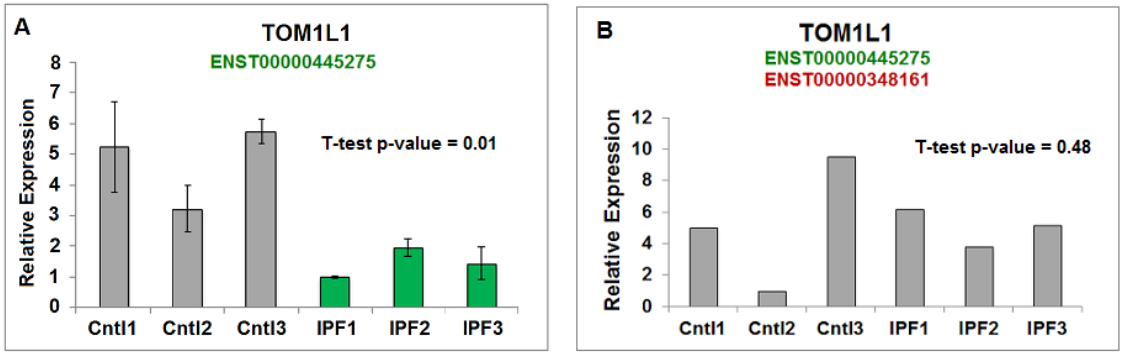


Figure 4.4: (A)The validation result for the down-regulated (in green color) isoform ENST0000445275 of gene TOM1L1 [23]. (B)The qRT-PCR validation result of the common region of transcripts in gene TOM1L1 [23].

The predicted splicing variant was validated by quantitative RT-PCR analysis as shown in Figure 4.4, with the bar chart representing the relative expression values among three samples in each condition. One splicing variant was confirmed for the case study of TOM1L1, and the experiment was performed in triplicate. In order to infer the up or down regulation of other splicing variants, we also quantified the common region of the transcripts in genes TOM1L1 qRT-PCR experiments. Due to

the limited amount of sample tissues, we performed single validation experiment on each individual sample.

For TOM1L1, the PCR primers were designed at the region of the skipped exon. As shown in Figure 4.4A, the qRT-PCR analysis confirms the down-regulated expression of transcript ENST00000445275 in samples from IPF patients in a statistically significant manner (T-test p-value of 0.01). We further quantified the common region of the transcripts ENST00000445275 and ENST00000348161 in each individual sample tissue, and the results demonstrate a non-significant change (T-test p-value of 0.48) in expression abundance (Figure 4.4B). Collectively, we confirmed that the transcript ENST00000348161 is up-regulated as predicted.

Biological Insight

Aged lung has a predisposition for disrepair and for lung fibrosis [108, 111]. Recently, it has been shown that significant DNA methylation differences that account for changes in gene expression are associated with specific age-related disorders, and one of these genes is TOM1L1. TOM1L1 is known to be recruited to the endosome and can subsequently recruit clathrin. In addition, it has been reported that TOM1L1 is a regulating adaptor bridging activated EGFR with the endocytic machinery for internalization of activated EGFR [71]. Taking together, we can speculate that TOM1L1 could potentially serve as a marker for lung aging and maybe as a marker for susceptibility to lung fibrogenesis.

4.2.7 Conclusion

Several array studies have been conducted to improve our understanding of the molecular processes involved in lung fibrogenesis, and to develop biomarkers.

However, most of these studies are based on differential expression analysis at the gene level through microarray platforms. This type of analysis is a powerful tool in identifying gene patterns and pathways associated with IPF [10, 15]. Splicing variants encoded by non-differentially expressed genes across conditions may play an important role in human IPF. Thus, by using RNA-Seq, we focused on detecting alternative splice variants from those non-differentially expressed genes, which have not been identified in previous pulmonary fibrosis microarray research. We applied abundance and variance filters at gene and isoform levels for detecting the most consistent splicing events in a conservative way. Our approach of joint analysis of differential expression and differential splicing appears to be useful in identifying splicing variants of IPF. Similar analysis approaches may also be applicable to deciphering the pathobiology of other life-threatening diseases.

Chapter 5: Conclusions and Future Works

Given the opportunity that RNA-Seq technologies provided for characterizing and better understanding human transcriptomes, computational approaches that utilize either read counts or read coverage signals have been extensively developed. These methods have helped biomedical researchers to extract useful information from large amounts of sequencing data from human tissues. Deep biological insights may be generated to understand the pathological consequences of diseased cell development and differentiation and to eventually identify potential biomarkers for human life-threatening diseases. In this dissertation, two algorithms and tools and a computational workflow using RNA-Seq were presented to analyze human transcriptomes between healthy and diseased conditions with a great emphasis on alternative splicing.

First, a read count-based Expectation-Maximization (EM) algorithm and tool, RAEM, was presented for estimating relative transcript proportions. By utilizing short reads aligned to exonic regions of each gene, RAEM constructs an observed cDNA fragment-compatible matrix to capture the relationship between short reads and all annotated transcript isoforms. Then, it employs an EM algorithm to infer the cDNA fragment-originating matrix with maximized likelihood and estimates the relative transcript isoform proportions iteratively. We applied RAEM to predict microRNA-155 targets at isoform-level.

Second, to identify and specify the spliced regions and the associated types of differential splicing events, we developed a read coverage-based algorithm and tool, called dSpliceType. dSpliceType utilizes sequential dependency of normalized base-wise read coverage signals and a change-point analysis, followed by a parametric statistical hypothesis test using Schwarz Information Criterion (SIC) to detect sig-

nificant differential splicing events in the form of five well-known types. We applied dSpliceType to detect differential splicing events from H1 (human embryonic stem cell) and H1 differentiated neuronal progenitor cultured cell lines.

Finally, a novel computational workflow was developed to jointly study genes with differential expression and differential splicing between healthy and diseased conditions. The computational workflow is employed to study a human idiopathic pulmonary fibrosis (IPF) lung disease. The genes are presented from two dimensions in terms of both differential expression and differential splicing. Some splicing variants from non-differentially expressed genes have been detected and biologically validated as potential biomarkers of human IPF disease.

Many possible future projects can be extended from the presented work:

- **Possible Extensions of RAEM** As introduced in the Chapter 2, RAEM estimates relative transcript proportions/abundances based on the assumption that the reads are sequenced uniformly at each nucleotide location along the transcript isoforms. However, short reads are more likely to be generated non-uniformly. Therefore, RAEM can be extended by using more sophisticated statistical models, such as a generalized Poisson model, and to use stochastic EM algorithm to overcome local optimal problem if the likelihood function is not concave. While the stochastic EM algorithm may need more computational efforts, the extension of RAEM can easily lend itself to parallelization by estimating relative transcript abundances of several genes simultaneously.

- **Possible Extensions of dSpliceType** The current command-line version of dSpliceType is focused on detecting the five most well-known types of differential splicing events based on a transcript annotation database. There are several ways to enhance dSpliceType to be more useful and robust. 1) dSpliceType

can be extended to detect novel differential splicing events by incorporating novel junctions when extracting candidate splicing events. 2) The framework of dSpliceType can also be extended and applied to detect other complicated splicing events. 3) dSpliceType can be incorporated into the SAMMate GUI software to easily allow biomedical researchers who lack computational skills to analyze their RNA-Seq data. 4) Since dSpliceType detects splicing events based on each candidate splicing event, it is rather straight forward to parallelize multiple detection procedures.

- **Classification Utilizing both Differential Expression and Differential Splicing** Machine learning methods have been widely used for classifying disease samples from healthy controls. A wide range of classification methods have been applied, such as K-Nearest Neighbor, Random Forest, Support Vector Machine and Logistic Regression. In many applications, differentially expressed genes have been selected as features to discriminate disease samples from healthy controls. As differential splicing is the mechanism-revealing feature of human diseases, it should be a better idea to utilize both differential expression genes and differential splicing genes to develop mechanistic and more effective classifiers.

RAEM has been encoded in an in-house software suite called SAMMate, which is freely available at `http://sammate.sourceforge.net/`. dSpliceType is freely available at `http://orleans.cs.wayne.edu/dSpliceType/`.

APPENDIX A: BIOLOGICAL TERM AND ABBREVIATION

3'UTR: 3' Untranslated Region. 3'UTR is in the end of a mRNA but not translated into proteins. 3' UTR may contain sequences that regulate translation efficiency, mRNA stability, and polyadenylation signals.

5'UTR: 5' Untranslated Region. 5'UTR is at the beginning of a mRNA but not translated into proteins.

A3SS: Alternative 3' Splice Site. A type of alternative splicing in which two splice sites are recognized at 3' end exon of an alternative splicing event.

A5SS: Alternative 5' Splice Site. A type of alternative splicing in which two splice sites are recognized at 5' end exon of an alternative splicing event.

AS: Alternative splicing. A regulated process during gene expression by which a single gene can produce multiple spliced mRNAs and proteins.

cDNA: complementary DNA. A form of DNA artificially synthesized from a messenger RNA template and used in genetic engineering to produce gene clones.

CPA: Change Point Analysis. An analytical method that attempts to find a point along a sequence of data point values where the characteristics or distribution of the values before and after the point are different.

Chromosome: An organized structure of DNA and protein found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences.

DEA: Differential Expression Analysis. It refers to use statistical testing to decide whether an observed difference in read counts of a gene/isoform across samples of two conditions is significant, and not due to random variation.

DSA: Differential Splicing Analysis. It refers to detect whether the difference in

the relative abundance of the expressed transcripts in a gene across samples of two conditions is significant.

EM algorithm: Expectation maximization algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables.

Exon: A sequence of DNA that codes information transcribed to mRNA and translated to protein.

Exon-exon junction: A sequence fragment spanning two exons. One end is mapped on the end of the first exon, and the other end is mapped on the beginning of the second exon.

FDR: False Discovery Rate is a statistical method used to adjust raw p-values in multiple hypothesis testing.

FPKM: Fragments Per Kilobase of exon/transcript model per Million mapped fragments.

Intron: A segment of a gene between exons without coding information for proteins. Introns are removed by RNA splicing process when producing mature mRNA.

IPF: Idiopathic Pulmonary Fibrosis. An interstitial human lung disease of unknown cause and high mortality rate.

Isoform: Any of several different forms of the same protein. Different forms of a protein may be produced from the same gene by alternative splicing.

microRNA/miRNA: microRNAs are single-stranded RNA molecules of about 21-23 nucleotides in length. They are non-coding RNAs, and their main function is to down regulate gene expression.

mRNA: Messenger RNA. A molecule of RNA encodes a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the ribosomes for protein synthesis.

MXE: Mutually Exclusive Exons. A type of alternative splicing in which two different individual exons are spliced out between two shared ending exons of two transcripts.

NGS: Next-Generation Sequencing. A group of new sequencing technologies that can rapidly sequence DNA or mRNA on the gigabase scale and generate millions of short reads.

RI: Retained Intron. A type of alternative splicing in which an intron remains in the mature mRNA transcript.

RNA-Seq: One type of NGS technologies. It sequences cDNA in order to get information about a sample's RNA content, which is quickly becoming a promising tool in the study of diseased human transcriptomes.

RPKM: Reads Per Kilobase of exon/transcript model per Million mapped reads.

SE: Skipped Exon. A type of alternative splicing in which a cassette exon and its flanking introns are spliced out of the transcript.

SIC: Schwarz Information Criterion. A likelihood function based statistical criterion for model selection among a finite set of models.

Transcript: A sequence of RNA produced by transcription.

Transcript quantification: It refers to quantify the transcript expression abundance.

Transcription and Translation: The process by which DNA is used as a template to create mRNA is called transcription. The mRNA then undergoes a further process called translation where the mRNA is used to synthesize proteins.

Transcriptome: The complete set of all transcripts, such as mRNAs, small RNAs and lincRNAs, in a cell or a population of cells.

APPENDIX B: LIST OF PUBLICATIONS

Journal Publications

1. **N. Deng**, C. Sanchez, J. Lasky and D. Zhu, "Detecting splicing variants in idiopathic pulmonary fibrosis from non-differentially expressed genes," *PLoS ONE*, vol. 8, no. 7, pp. e68352, 2013.

2. **N. Deng**, A. Puetter, K. Zhang, K. Johnson, Z. Zhao, C. Taylor, EK. Flemington and D. Zhu, "Isoform-level microRNA-155 Target Prediction using RNA-seq," *Nucleic Acids Research*, vol. 39, no. 9, pp. e61–e61, 2011.

3. G. Xu, **N. Deng**, Z. Zhao, EK. Flemington and D. Zhu, "SAMMate: A GUI tool for processing short read alignment information in SAM/BAM format," *Source Code for Biology and Medicine*, vol. 6, no. 1, pp. 2, 2011.

4. G. Xu, C. Fewll, C. Taylor, **N. Deng**, D. Hedges, X. Wang, K. Zhang, M. Lacey, H. Zhang, Q. Yin, J. Cameron, Z. Lin, D. Zhu and EK. Flemington, "Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq," *RNA*, vol. 16, no. 8, pp. 1610–22, 2010.

5. L. Zhang, G. Xu, **N. Deng**, C. Taylor, D. Zhu and EK. Flemington, "Quantitative and Qualitative RNA-Seq-Based Evaluation of Epstein-Barr virus Transcription in Type I Latency Burkitt's Lymphoma Cells," *Journal of Virology*, vol. 84, no. 24, pp. 13053–58, 2010.

Conference Publications

1. **N. Deng** and D. Zhu, "dSpliceType: a multivariate model for detecting various types of differential splicing events using RNA-Seq," ISBRA, 2014, Conference Paper, page 333-344.

2. **N. Deng** and D. Zhu, "Detecting various types of differential splicing events using RNA-Seq data," ACM, 2013, Conference Paper, page 124-132.

3. Z. Zhao, T. Nguyen, **N. Deng**, K. Johnson and D. Zhu, "SPATA: A seeding and patching algorithm for de novo transcriptome assembly," IEEE, 2011, Workshop Paper, 10.1109/BIBMW.2011.6112351.

4. T. Nguyen, **N. Deng**, G. Xu, Z. Duan and D. Zhu, "iQuant: A fast yet accurate GUI tool for transcript quantification," IEEE, 2011, Workshop Paper, 10.1109/BIBMW.2011.611256.

APPENDIX C: COPYRIGHTS

Various copyright/licensing agreements allowing the use of previously published material is presented in this appendix.

Publication Rights Policies

What is our policy?

For the majority of journals[1] published by Oxford University Press, we have a policy of acquiring a sole and exclusive licence for all published content, rather than asking authors to transfer ownership of their copyright, which has been common practice in the past. We believe this policy more carefully balances the interests of our authors with our need to maintain the viability and reputation of the journals through which our authors are accorded status, recognition and widespread distribution. In developing this policy we have been guided by the following principles:

- As a university press and not-for-profit academic publisher, we rely heavily on the good relationships we have with our authors. Having a licensing policy which enables an author to be identified as the owner of the copyright in an article is one of the key ways of demonstrating how highly we value these relationships.
- An exclusive licence enables the centralised and efficient management of permissions and licencing, ensuring the widest dissemination of the content through intermediaries;
- Exclusive rights also enable OUP to take measures on behalf of our authors against infringement, inappropriate use of an article, libel or plagiarism;
- At the same time, by maintaining exclusive rights, in all media for all published content, we can monitor and uphold the integrity of an article once refereed and accepted for publication to be maintained;

Where to get a copy of the Licence to Publish

OUP cannot publish your article until a completed licence form has been received. You should receive a form as soon as your article is accepted for publication.

Footnotes to this section

1. A small number of OUP Journals still have a policy of requesting a full Assignment of Copyright. If unclear about the policy of the Journal concerned, please contact the Editorial office to clarify.

Government employees

- If you are or were a UK Crown servant and the article has been written in that capacity, we have an arrangement with HMSO to enable us to publish it while acknowledging that it is Crown Copyright. Please inform the Editorial office or Oxford University Press at the time of acceptance or as soon as possible that the article is Crown Copyright, so that we can ensure the appropriate acknowledgement and copyright line are used, as required by our arrangement with HMSO.
- If you are a US Government employee and the article has been written in that capacity, we acknowledge that the Licence to Publish applies only to the extent allowable by US law.

Re-use of third party content as part of your Oxford Journals article

- As part of your article, you may wish to reuse material sourced from third parties such as other publishers, authors, museums, art galleries etc. To assist with this process, we have a Permission Request form and accompanying Guidelines that specifies the rights required in order for third party material to be published as part of your Article. For a copy of this form, please email.
- Responsibility for clearing these third party permissions must be borne by the Author, and this process completed as soon as possible - preferably before acceptance of the manuscript, but if not possible, before the Article reaches the Production stage of the process.

Rights retained by ALL Oxford Journal Authors

- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

The only exception to this is for the re-use of material for commercial purposes, as defined in the information available via the above url. Permission for this kind of re-use is required and can be obtained by using Rightslink:

With Copyright Clearance Center's Rightslink ® service it's faster and easier than ever before to secure permission from OUP titles to be republished in a coursepack, book, CD-ROM/DVD, brochure or pamphlet, journal or magazine, newsletter, newspaper, make a photocopy, or translate.

- Simply visit: www.oxfordjournals.org and locate your desired content.
- Click on (Order Permissions) within the table of contents and/ or at the bottom article's abstract to open the following page:
- Select the way you would like to reuse the content
- Create an account or login to your existing account
- Accept the terms and conditions and permission is granted

For questions about using the Rightslink service, please contact Customer Support via phone 877/622-5543 (toll free) or 978/777-9929, or email Rightslink customer care.

Preprint use of Oxford Journals content

- For the majority of Oxford Journals, prior to acceptance for publication, authors retain the right to make a pre-print [*A preprint is defined here as un-refereed author version of the article*] version of the article available on your own personal website and/or that of your employer and/or in free public servers of preprints and/or articles in your subject area, provided that where possible.
    - You acknowledge that the article has been accepted for publication in [Journal Title] ©: [year] [owner as specified on the article] Published by Oxford University Press [on behalf of xxxxxx]. All rights reserved.
    - Once the article has been published, we do not require that preprint versions are removed from where they are available. However, we do ask that these are not updated or replaced with the finally published version. Once an article is published, a link could be provided to the final authoritative version on the Oxford Journals Web site. Where possible, the preprint notice should be amended to:
    - This is an electronic version of an article published in [include the complete citation information for the final version of the Article as published in the print edition of the Journal.]
- Once an article is accepted for publication, an author may not make a pre-print available as above or replace an existing pre-print with the final published version. **NB**There are some Oxford Journals such as the Journal of the National Cancer Institute, which do not permit any kind of preprint use. For clarification of the preprint policy for any journal please contact the Rights and New Business Development Department.

Postprint use of Oxford Journals content:

[*A postprint is defined here as being the final draft author manuscript as accepted for publication, following peer review, BUT before it has undergone the copyediting and proof correction process*].

We have detailed policies on the use of postprints for all of our journals. To view these for individual journals please refer to the author self archiving policies on journal homepages. If you require further information please contact the Rights and New Business Development Department.

Other uses by authors should be authorized by Oxford Journals through the Rights and New Business Development Department.

Additional Rights retained by the Author when publishing in an Oxford Open participating journal

Please note that these rights only apply to content published in an Oxford Journal on an Open Access basis in exchange for payment of an author charge. For more details about how Oxford Open works please click here.

The right to reproduce, disseminate or display articles published under this model for educational purposes, provided that:

- the original authorship is properly and fully attributed;
- the Journal and OUP are attributed as the original place of publication with the correct citation details given;
- if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated
- the right to deposit the postprint and/or URL or PDF of the finally published version of the article into an institutional or centrally organized repository, immediately upon publication

Commercial Use of Open Access version

For all articles published under a Creative Commons Attribution Licence (CC BY 3.0) or an Open Government Licence permission is not required to make any kind of commercial use of the material.

For all articles published under a Creative Commons Non-Commercial Attribution Licence (CC BY NC 3.0) or a Non-Commercial Government Licence permission is required for all commercial reuse. In order to request permission please contact the Rights and New Business Development department: you want to use and a brief description of the intended use.

Commercial re-use guidelines for open access content

Definition of commercial use: any re-use of material from the Open Access part of an Oxford Journal for the commercial gain of the user and/or their employing institution. In particular,

- re-use by a non-author/third party/other publisher of parts of or all of an article or articles in another publication (journal or book) to be sold for commercial purposes. Permission to reproduce selected figures will generally be granted free of charge, although OUP reserves the right to levy a fee for the use of these and/or the full text of an article/articles
- the proactive supply of multiple print or electronic copies of items taken from the Journal to third parties on a systematic basis for marketing purposes. Permission for this kind of reuse should be obtained from the publisher, who retains the right to levy an appropriate fee
- re-use by an author of parts of or all of an article in other publications from commercial organizations. Permission for this kind of reuse should be obtained from the publisher. We would consider this to be commercial reuse but would not normally charge a permission fee if the author is involved.

NB: Please note that any income generated from permissions granted for this kind of use will be returned directly to the journal itself in order to help minimise the costs of making content from it available on an Open Access basis.

Permissions

- All requests to reuse the article, in whole or in part, in another publication will be handled by Oxford Journals. Unless otherwise stated, any permission fees will be retained by the Journal concerned. Where possible, any requests to reproduce substantial parts of the article (including in other Oxford University Press publications) will be subject to your approval (which is deemed to be given if we have not heard from you within 4 weeks of the permission being granted).

Open-Access License

No Permission Required

ACM Information for Authors

ACM Author Rights

ACM exists to support the needs of the computing community. For over sixty years ACM has developed publications and publication policies to maximize the visibility, impact, and reach of the research it publishes to a global community of researchers, educators, students, and practitioners. ACM has achieved its high impact, high quality, widely-read portfolio of publications with:

- Affordably priced publications
- Liberal Author rights policies
- Wide-spread, perpetual access to ACM publications via a leading-edge technology platform
- Sustainability of the good work of ACM that benefits the profession

Choose

Authors have the option to choose the level of rights management they prefer. ACM offers three different options for authors to manage the publication rights to their work.

- Authors who want ACM to manage the rights and permissions associated with their work, which includes defending against improper use by third parties, can use ACM's traditional copyright transfer agreement.
- Authors who prefer to retain copyright of their work can sign an exclusive licensing agreement, which gives ACM the right but not the obligation to defend the work against improper use by third parties.
- Authors who wish to retain all rights to their work can choose ACM's author-pays option, which allows for perpetual open access through the ACM Digital Library. Authors choosing the author-pays option can give ACM non-exclusive permission to publish, sign ACM's exclusive licensing agreement or sign ACM's traditional copyright transfer agreement.

Post

Authors can post the accepted, peer-reviewed version prepared by the author—known as the "pre-print"—to the following sites, with a DOI pointer to the Definitive Version in the ACM Digital Library.

- On Author's own Home Page *and*
- On Author's Institutional Repository *and*
- In any repository legally mandated by the agency funding the research on which the work is based.
- Prior to submission to ACM for peer-review, authors may post their original work in any informal, non-peer-reviewed aggregation or collection.

Distribute

Authors can post an *Author-Izer* link enabling free downloads of the Definitive Version of the work permanently maintained in the ACM Digital Library

- On the Author's own Home Page *or*
- In the Author's Institutional Repository.

Reuse

Authors can reuse any portion of their own work in a new work of *their own* (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is *not* the editor, requires permission and usually a republication fee.

Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).

- Commercially produced course-packs that are *sold* to students require permission and possibly a fee.

Create

ACM's copyright and publishing license include the right to make Derivative Works or new versions. For example, translations are "Derivative Works." By copyright or license, ACM may have its publications translated. However, ACM Authors continue to hold perpetual rights to revise their own works without seeking permission from ACM.

- If the revision is minor, i.e., less than 25% of new substantive material, then the work should still have ACM's publishing notice, DOI pointer to the Definitive Version, and be labeled a "Minor Revision of"
- If the revision is major, i.e., 25% or more of new substantive material, then ACM considers this a new work in which the author retains full copyright ownership (despite ACM's copyright or license in the original published article) and the author need only cite the work from which this new one is derived.

Minor Revisions and Updates to works already published in the ACM Digital Library are welcomed with the approval of the appropriate Editor-in-Chief or Program Chair.

Retain

Authors retain all *perpetual rights* laid out in the ACM Author Rights and Publishing Policy, including, but not limited to:

- Sole ownership and control of third-party permissions to use for artistic images intended for exploitation in other contexts
- All patent and moral rights
- Ownership and control of third-party permissions to use of software published by ACM

Copyright of Springer

COPYRIGHT ACT

In most countries of the world, authors enjoy protection of their intellectual property that appears in books, journal articles and parts thereof, such as illustrations, plans, tables and animations. Protected works include literary and scientific works, such as writings, speeches and computer programs. Only personal intellectual creations are protected.
The person who writes one of the aforementioned works is defined as the creator/author. Co-authorship applies if two or more persons create a work together.
Notice of Copyright is printed in general on the verso of the title page of a book or on the header or footer of a journal article. Notice of Copyright provides information regarding the date of first publication of the work and the holder of copyright. Proper notice of copyright helps to protect the integrity of the work and to fight copyright infringement.

CONTENTS OF COPYRIGHT

Moral Rights cover an author's authority to decide whether his work should be published and whether the published work should bear the author's name.
Exploitation Rights entitle an author to decide whether copies of the work should be reproduced (Right of Reproduction) and whether these copies should be offered to the public (Right of Distribution). Right of Reproduction is the right to make copies of the work, irrespective of method or number. Right of Distribution is the right to offer to the public the aforementioned produced copies.

COPYRIGHT LICENCES

Author are free to publish their work by themselves or transfer the exploitation rights to a publisher; e.g. Springer. In order to be entitled to make use of these rights, the publisher asks the author to sign a publishing agreement granting the publisher the sole right to reproduce, publish, distribute and make available to the public the work in print and electronic format. Authors and the publisher should always define their relationship in a publishing agreement. Springer offers a large variety of such contracts for all kind of works. Authors should contact their Springer publishing editor for more details.
Prerequisite of the transfer of exclusive publishing rights is that the author has not already signed such rights to third parties (e.g. another publisher) and that the work has not heretofore been published in whole or in part.

Consequence of having granted exclusive rights to Springer indicates also that an author agrees not to release with another publisher any publication similar to the work published with Springer.
Authors retain, in addition to uses permitted by law (e.g. U.S. Copyright Law, Section 107, Fair Use; German Copyright Act, Section 51, Fair Dealing) the right to communicate the content of the work to other scientists, to share the work with them in manuscript form, to perform or present the work or to use the content for non-commercial internal and educational purposes.

LIMITATIONS ON COPYRIGHT

To the extent required by the purpose, it is permissible to reproduce, distribute and publicly communicate single works that have already been published, included in an independent scientific work in order to clarify their contents. The limits of fair dealing will vary according to special circumstances. Acknowledgement needs to be given to the original source of publication. Omission of a sufficient acknowledgement constitutes an infringement of the copyright of the cited work.
Under certain circumstances, it is permissible to make single copies of a work for private, non-commercial use; e.g. for personal scientific use or for teaching in non-commercial institutions of education. These copies may be neither disseminated nor used for public communication.

DURATION OF COPYRIGHT

Copyright is legally valid for a fixed period of time. The length of the period varies depending on the copyright laws of each country. It is usually from 50 to 70 years after the death of the author.
Once this term has expired, however, legal rights to the work also expire. After that, the work becomes part of the public domain and can be used freely.

RELATED RIGHTS

Scientific Editions which consist of non-copyrighted works (i.e. public domain works) are protected by copyright if they represent the result of scientific analysis and differ in significant manner from previous editions of the works. Copyright protection expires 25 years after publication of the scientific edition.
Photographs are also protected by copyright. Copyright protection expires 50 years after the publication of the photograph.

INHERITANCE OF COPYRIGHT

Copyright may be transmitted by inheritance. The author’s legal successor shall have the rights enjoyed by the deceased author according to the arrangements of local copyright laws.

INFRINGEMENT OF COPYRIGHT

Copyright is protected both domestically and internationally according to the laws and treaties of each nation. Nevertheless, copyright infringements often do occur.
Springer takes care of an author’s right and undertakes any necessary steps to protect these rights against infringement by third parties.
Any person or legal entity that infringes on the copyright of a Springer author will be urged to cease and desist from the wrongdoing and provide detailed information about the infringement.
Moreover, destruction of all copies unlawfully manufactured and distributed will be required.

REFERENCES

[1] M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, and J. C. Venter. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature genetics*, 4(4):373–380, 1993.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, and K. Roberts. *Molecular Biology of the Cell, 4th Edition*. National Center for Biotechnology InformationÕs Bookshelf, 2011.

[3] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[4] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 2012.

[5] M. Aschoff, A. Hotz-Wagenblatt, K.-H. Glatting, M. Fischer, R. Eils, and R. König. Splicingcompass: differential splicing detection using RNA-Seq data. *Bioinformatics*, 29(9):1141–1148, 2013.

[6] W. B. Barbazuk, S. J. Emrich, H. D. Chen, L. Li, and P. S. Schnable. SNP discovery via 454 transcriptome sequencing. *The plant journal*, 51(5):910–918, 2007.

[7] J. Beane, J. Vick, F. Schembri, C. Anderlind, A. Gower, J. Campbell, L. Luo, X. H. Zhang, J. Xiao, Y. O. Alekseyev, et al. Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prevention Research*, 4(6):803–817, 2011.

[8] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[9] R. Bohnert and G. Rätsch. rQuant. web: a tool for RNA-Seq-based transcript quantitation. *Nucleic acids research*, 38(suppl 2):W348–W351, 2010.

[10] K. Boon, N. W. Bailey, J. Yang, M. P. Steel, S. Groshong, D. Kervitsky, K. K. Brown, M. I. Schwarz, and D. A. Schwartz. Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (IPF). *PLoS One*, 4(4):e5134, 2009.

[11] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–634, 2000.

[12] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. 1994.

[13] J. Chen. *Parametric statistical change point analysis*. Birkhauser Boston, 2012.

[14] J. Chen and Y.-P. Wang. A statistical change point model approach for the detection of DNA copy number variations in array CGH data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(4):529–541, 2009.

[15] J.-H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, D. Dakhlallah, J. Price, D. Bornman, S. Zhang, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC medical genomics*, 4(1):8, 2011.

[16] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–619, 2008.

[17] T. A. Cooper, L. Wan, and G. Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, 2009.

[18] W. R. Coward, G. Saini, and G. Jenkins. The pathogenesis of idiopathic pulmonary fibrosis. *Therapeutic advances in respiratory disease*, 4(6):367–388, 2010.

[19] C. J. Creighton, A. K. Nagaraja, S. M. Hanash, M. M. Matzuk, and P. H. Gunaratne. A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *Rna*, 14(11):2290–2296, 2008.

[20] F. Crick et al. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, 1970.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[22] 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*, 39(9):e61–e61, 2011.

[23] 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*, 8(7):e68352, 2013.

[24] 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*, page 124. ACM, 2013.

[25] N. Deng and D. Zhu. dSpliceType: A multivariate model for detecting various types of differential splicing events using RNA-Seq. *Bioinformatics Research and Applications, Lecture Notes in Computer Science*, 8492:322–333, 2014.

[26] F. Denoeud, J.-M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, et al. Annotating genomes with massive-scale RNA sequencing. *Genome Biol*, 9(12):R175, 2008.

[27] M. L. Eaton. *Multivariate statistics: a vector space approach*. Wiley New York, 1983.

[28] G. Edwalds-Gilbert. Regulation of mRNA Splicing by Signal Transduction. *Nature Education*, 2010.

[29] A. El-Karef, M. Kaito, H. Tanaka, K. Ikeda, T. Nishioka, N. Fujita, H. Inada, Y. Adachi, N. Kawada, Y. Nakajima, et al. Expression of large tenascin-C splice variants by hepatic stellate cells/myofibroblasts in chronic hepatitis C. *Journal of hepatology*, 46(4):664–673, 2007.

[30] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, D. S. Marks, et al. MicroRNA targets in drosophila. *Genome biology*, 5(1):R1–R1, 2004.

[31] J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. In *Research in Computational Molecular Biology*, pages 138–157. Springer, 2010.

[32] J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. *Journal of Computational Biology*, 18(3):305–321, 2011.

[33] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

[34] P. J. Gardina, T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7(1):325, 2006.

[35] D. S. Gerhard, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, et al. The status, quality,

and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (mgc). *Genome research*, 14(10B):2121–2127, 2004.

[36] E. M. Glare, M. Divjak, J. M. Rolland, and E. H. Walters. Asthmatic airway biopsy specimens are more likely to express the IL-4 alternative splice variant IL-4δ2. *Journal of allergy and clinical immunology*, 104(5):978–982, 1999.

[37] M. Gonzàlez-Porta, M. Calvo, M. Sammeth, and R. Guigó. Estimation of alternative splicing variability in human populations. *Genome research*, 22(3):528–538, 2012.

[38] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[39] 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*, 40(20):10073–10083, 2012.

[40] M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, R. Corbett, M. J. Tang, Y.-C. Hou, T. J. Pugh, et al. Alternative expression analysis by RNA sequencing. *Nature Methods*, 7(10):843–847, 2010.

[41] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic acids research*, 36(suppl 1):D154–D158, 2008.

[42] P. K. Gupta. Single-molecule DNA sequencing technologies for future genomics research. *Trends in biotechnology*, 26(11):602–611, 2008.

[43] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction

of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5):503–510, 2010.

[44] K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131, 2010.

[45] K. Honda, T. Yamada, M. Seike, Y. Hayashida, M. Idogawa, T. Kondo, Y. Ino, and S. Hirohashi. Alternative splice variant of actinin-4 in small cell lung cancer. *Oncogene*, 23(30):5257–5262, 2004.

[46] 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*, 41(2):e39–e39, 2013.

[47] J. C. Huang, T. Babak, T. W. Corson, G. Chua, S. Khan, B. L. Gallie, T. R. Hughes, B. J. Blencowe, B. J. Frey, and Q. D. Morris. Using expression profiling data to identify human microRNA targets. *Nature methods*, 4(12):1045–1049, 2007.

[48] W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16):1963–1970, 2006.

[49] J. L. Jensen. *Statistics for petroleum engineers and geoscientists*, volume 2. Access Online via Elsevier, 2000.

[50] H. Jiang and W. H. Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

[51] 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*, 7(12):1009–1015, 2010.

[52] H. Keren, G. Lev-Maor, and G. Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

[53] 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*, 14(4):R36, 2013.

[54] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, and B.-T. Zhang. miTarget: microRNA target gene prediction using a support vector machine. *BMC bioinformatics*, 7(1):411, 2006.

[55] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, et al. CAGE: cap analysis of gene expression. *Nature methods*, 3(3):211–222, 2006.

[56] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, et al. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, 2005.

[57] T. Kwan, D. Benovoy, C. Dias, S. Gurd, D. Serre, H. Zuzan, T. A. Clark, A. Schweitzer, M. K. Staples, H. Wang, et al. Heritability of alternative splicing in the human genome. *Genome Research*, 17(8):1210–1218, 2007.

[58] E. Laajala, T. Aittokallio, R. Lahesmaa, and L. L. Elo. Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. *Genome Biol*, 10(7):R77, 2009.

[59] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[60] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[61] I. Lee, S. S. Ajay, J. I. Yook, H. S. Kim, S. H. Hong, N. H. Kim, S. M. Dhanasekaran, A. M. Chinnaiyan, and B. D. Athey. New class of microRNA targets containing simultaneous 5-UTR and 3-UTR interaction sites. *Genome research*, 19(7):1175–1183, 2009.

[62] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[63] B. Li and C. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[64] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[65] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[66] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[67] J. Li, H. Jiang, and W. Wong. Method modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol*, 11(5):R25, 2010.

[68] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[69] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[70] H.-X. Liu, L. Cartegni, M. Q. Zhang, and A. R. Krainer. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature Genetics*, 27(1):55–58, 2001.

[71] N. S. Liu, L. S. Loo, E. Loh, L.-F. Seet, and W. Hong. Participation of Tom1L1 in EGF-stimulated endocytosis of EGF receptor. *The EMBO journal*, 28(22):3485–3499, 2009.

[72] D. A. Lynch, J. D. Godwin, S. Safrin, K. M. Starko, P. Hormel, K. K. Brown, G. Raghu, T. E. King Jr, W. Z. Bradford, D. A. Schwartz, et al. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *American journal of respiratory and critical care medicine*, 172(4):488–493, 2005.

[73] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, 2009.

[74] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic acids research*, 37(suppl 2):W273–W276, 2009.

[75] A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, 2005.

[76] M. L. Metzker. Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009.

[77] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra. Profiling the HeLa S3

transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45(1):81, 2008.

[78] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

[79] A. F. Muro, F. A. Moretti, B. B. Moore, M. Yan, R. G. Atrasz, C. A. Wilke, K. R. Flaherty, F. J. Martinez, J. L. Tsui, D. Sheppard, et al. An essential role for fibronectin extra type III domain A in pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 177(6):638, 2008.

[80] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[81] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[82] T. C. Nguyen, N. Deng, G. Xu, Z. Duan, and D. Zhu. iquant: A fast yet accurate gui tool for transcript quantification. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pages 1048–1050. IEEE, 2011.

[83] T. C. Nguyen, N. Deng, and D. Zhu. SASeq: A Selective and Adaptive Shrinkage Approach to Identify and Quantify Condition-Specific Transcripts using RNA-Seq. *arXiv preprint arXiv:1208.3619*, 2012.

[84] M. Nicolae, S. Mangul, I. I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, 6(1):9, 2011.

[85] P. Nicolas, A. Leduc, S. Robin, S. Rasmussen, H. Jarmer, and P. Bessieres. Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics*, 25(18):2341–2347, 2009.

[86] L. Pachter. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889*, 2011.

[87] 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*, 40(12):1413–1415, 2008.

[88] C. Quince, A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods*, 6(9):639–641, 2009.

[89] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[90] N. Rajewsky. microRNA target predictions in animals. *Nature genetics*, 38:S8–S13, 2006.

[91] A. K. Ramani, J. A. Calarco, Q. Pan, S. Mavandadi, Y. Wang, A. C. Nelson, L. J. Lee, Q. Morris, B. J. Blencowe, M. Zhen, et al. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome research*, 21(2):342–348, 2011.

[92] H. Richard, M. H. Schulz, M. Sultan, A. Nürnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic acids research*, 38(10):e112–e112, 2010.

[93] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, L. Pachter, et al. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3):R22, 2011.

[94] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[95] J. Salzman, H. Jiang, and W. H. Wong. Statistical modeling of RNA-Seq data. *Statistical Science*, 26(1):62–83, 2011.

[96] M. Sammeth, S. Foissac, and R. Guigo. A general definition and nomenclature for alternative splicing events. *PLoS computational biology*, 4(8):e1000147, 2008.

[97] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[98] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[99] G. T. SEAH, P. S. GAO, J. M. HOPKIN, and G. A. ROOK. Interleukin-4 and its alternatively spliced variant (IL-4 δ 2) in patients with atopic asthma. *American journal of respiratory and critical care medicine*, 164(6):1016–1018, 2001.

[100] 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*, 40(8):e61–e61, 2012.

[101] S. Shen, C. C. Warzecha, R. P. Carstens, and Y. Xing. MADS+: discovery of differential splicing events from Affymetrix exon junction array data. *Bioinformatics*, 26(2):268–269, 2010.

[102] 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*, 27(19):2633–2640, 2011.

[103] A. D. Smith, Z. Xuan, and M. Q. Zhang. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics*, 9(1):128, 2008.

[104] D. R. Smith, A. R. Quinlan, H. E. Peckham, K. Makowsky, W. Tao, B. Woolf, L. Shen, W. F. Donahue, N. Tusneem, M. P. Stromberg, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome research*, 18(10):1638–1642, 2008.

[105] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[106] S. Srivastava and L. Chen. A two-parameter generalized poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38(17):e170–e170, 2010.

[107] O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential RNA-transcript expression from read counts. *Nat Precedings*, 2010.

[108] V. Sueblinvong, D. C. Neujahr, S. T. Mills, S. Roser-Page, J. D. Ritzenthaler, D. Guidot, M. Rojas, and J. Roman. Predisposition for disrepair in the aged lung. *The American journal of the medical sciences*, 344(1):41, 2012.

[109] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, 2008.

[110] H. Tanaka, A. El-Karef, M. Kaito, N. Kinoshita, N. Fujita, S. Horiike, S. Watanabe, T. Yoshida, and Y. Adachi. Circulating level of large splice variants of tenascin-C is a marker of piecemeal necrosis activity in patients with chronic hepatitis c. *Liver International*, 26(3):311–318, 2006.

[111] E. Torres-González, M. Bueno, A. Tanaka, L. T. Krug, D.-S. Cheng, V. V. Polosukhin, D. Sorescu, W. E. Lawson, T. S. Blackwell, M. Rojas, et al. Role of endoplasmic reticulum stress in age-related susceptibility to lung fibrosis. *American journal of respiratory cell and molecular biology*, 46(6):748–756, 2012.

[112] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[113] 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*, 7(3):562–578, 2012.

[114] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[115] 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*, 98(9):5116–5121, 2001.

[116] V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler, et al. Serial analysis of gene expression. *Science-AAAS-Weekly Paper Edition*, 270(5235):484–486, 1995.

[117] J. P. Venables. Aberrant and alternative splicing in cancer. *Cancer Research*, 64(21):7647–7654, 2004.

[118] 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*, 456(7221):470–476, 2008.

[119] G.-S. Wang and T. A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

[120] X. Wang and X. Wang. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic acids research*, 34(5):1646–1652, 2006.

[121] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[122] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

[123] 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*, 27(21):3010–3016, 2011.

[124] Z. Wu, X. Wang, and X. Zhang. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, 27(4):502–508, 2011.

[125] L. Xi, A. Feber, V. Gupta, M. Wu, A. D. Bergemann, R. J. Landreneau, V. R. Litle, A. Pennathur, J. D. Luketich, and T. E. Godfrey. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Research*, 36(20):6535–6547, 2008.

[126] 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*, 14(8):1470–1479, 2008.

[127] Y. Xing, T. Yu, Y. N. Wu, M. Roy, J. Kim, and C. Lee. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic acids research*, 34(10):3150–3160, 2006.

[128] G. Xu, N. Deng, Z. Zhao, T. Judeh, E. Flemington, and D. Zhu. SAMMate: a GUI tool for processing short read alignments in SAM/BAM format. *Source code for biology and medicine*, 6(1):2, 2011.

[129] G. Xu, C. Fewell, C. Taylor, N. Deng, D. Hedges, X. Wang, K. Zhang, M. Lacey, H. Zhang, Q. Yin, et al. Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *Rna*, 16(8):1610–1622, 2010.

[130] G. W. Yeo, X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal, and F. H. Gage. Alternative splicing events identified in human embryonic stem cells and neural progenitors. *PLoS Computational Biology*, 3(10):e196, 2007.

[131] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe. Naive Bayes for microRNA target predictions-machine learning for microRNA targets. *Bioinformatics*, 23(22):2987–2992, 2007.

[132] D. Yue, H. Liu, and Y. Huang. Survey of computational algorithms for microRNA target prediction. *Current genomics*, 10(7):478, 2009.

ABSTRACT

ALGORITHMS AND TOOLS FOR COMPUTATIONAL ANALYSIS OF HUMAN TRANSCRIPTOME USING RNA-SEQ

by

NAN DENG

December 2014

Advisor: Dr. Dongxiao Zhu

Major: Computer Science

Degree: Doctor of Philosophy

Alternative splicing plays a key role in regulating gene expression, and more than 90% of human genes are alternatively spliced through different types of alternative splicing. Dysregulated alternative splicing events have been linked to a number of human diseases. Recently, high-throughput RNA-Seq technologies have provided unprecedented opportunities to better characterize and understand transcriptomes, in particular useful for the detection of splicing variants between healthy and diseased human transcriptomes.

We have developed two novel algorithms and tools and a computational workflow to interrogate human transcriptomes between healthy and diseased conditions. The first is a read count-based Expectation-Maximization (EM) algorithm and tool, which is called RAEM. It estimates relative transcript isoform proportions by maximizing the likelihood in each gene. The RAEM algorithm has been encoded in our published software suite, SAMMate. We have employed RAEM to predict isoform-level microRNA-155 targets. The second is called dSpliceType, which is a read coverage-based algorithm and tool to detect differential splicing events. It utilizes sequential dependency of normalized base-wise read coverage signals and a change-point

analysis, followed by a parametric statistical hypothesis test using Schwarz Information Criterion (SIC) to detect significant differential splicing events in the form of the five well-known splicing types. The results of both simulation and real-world studies demonstrate that dSpliceType is an efficient computational tool for detecting various types of differential splicing events from a wide range of expressed genes. Finally, we developed a novel computational workflow to jointly study human diseases in terms of both differential expression and differential splicing. The workflow has been used to detect differential splicing variants from non-differentially expressed genes of human idiopathic pulmonary fibrosis (IPF) lung disease.

AUTOBIOGRAPHICAL STATEMENT

Nan Deng was born in Beijing, P.R.China. She received her Bachelor of Engineering in Computer Science from Beijing University of Posts and Telecommunications (BUPT) in 1998, and received her Master of Art in Geographic Information Sciences for Development and Environment (GISDE) from Clark University in 2008. She joined the University of New Orleans in the Fall of 2009 to pursue her Ph.D. in Computer Science. She transferred with her advisor Dr. Dongxiao Zhu to Wayne State University continuing her Ph.D. studies from the Fall of 2011. Her research interests include developing novel algorithms and tools for transcriptome characterization and identification using RNA-Seq, in particular detecting various types of differential splicing events between healthy and diseased human transcriptomes.