University of Kentucky
UKnowledge

Theses and Dissertations--Computer Science | Computer Science

2013

A NOVEL COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ANALYSIS WITH RNA-SEQ DATA

Yin Hu
University of Kentucky, snowy8677@gmail.com

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Hu, Yin, "A NOVEL COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ANALYSIS WITH RNA-SEQ DATA" (2013). *Theses and Dissertations--Computer Science*. 17.
https://uknowledge.uky.edu/cs_etds/17

This Doctoral Dissertation is brought to you for free and open access by the Computer Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Computer Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Yin Hu, Student

Dr. Jinze Liu, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

A NOVEL COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ANALYSIS WITH RNA-SEQ DATA

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Yin Hu
Lexington, Kentucky

Director: Dr. Jinze Liu, Associate Professor of Computer Science
Lexington, Kentucky 2013

Copyright © Yin Hu 2013

ABSTRACT OF DISSERTATION

A NOVEL COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ANALYSIS WITH RNA-SEQ DATA

The advance of high-throughput sequencing technologies and their application on mRNA transcriptome sequencing (RNA-seq) have enabled comprehensive and unbiased profiling of the landscape of transcription in a cell. In order to address the current limitation of analyzing accuracy and scalability in transcriptome analysis, a novel computational framework has been developed on large-scale RNA-seq datasets with no dependence on transcript annotations. Directly from raw reads, a probabilistic approach is first applied to infer the best transcript fragment alignments from paired-end reads. Empowered by the identification of alternative splicing modules, this framework then performs precise and efficient differential analysis at automatically detected alternative splicing variants, which circumvents the need of full transcript reconstruction and quantification. Beyond the scope of classical group-wise analysis, a clustering scheme is further described for mining prominent consistency among samples in transcription, breaking the restriction of presumed grouping. The performance of the framework has been demonstrated by a series of simulation studies and real datasets, including the Cancer Genome Atlas (TCGA) breast cancer analysis. The successful applications have suggested the unprecedented opportunity in using differential transcription analysis to reveal variations in the mRNA transcriptome in response to cellular differentiation or effects of diseases.

KEYWORDS: RNA-seq, transcriptome, algorithm, statistical inference, data mining

Author's signature: Yin Hu

Date: December 20, 2013

A NOVEL COMPUTATIONAL FRAMEWORK FOR TRANSCRIPTOME ANALYSIS WITH RNA-SEQ DATA

By
Yin Hu

Director of Dissertation: Jinze Liu

Director of Graduate Studies: Miroslaw Truszczyński

Date: December 20, 2013

ACKNOWLEDGMENTS

Pursing a Ph.D is a both painful and enjoyable experience. Through this challenging journey, I encountered frustration in perplexity and I tasted happiness in success. Now standing at the terminal, I realized I could never get there without the help from many people. I would like to express my sincere gratitude to all these people.

My first and foremost gratitude must go to my advisor, Dr. Jinze Liu for continuous support of my Ph.D study. Her enthusiasm, motivation and patience inspired my interest for research and guided me proceeding towards the right direction. She encourages me to think as an independent scientist which is consistent with my long-term career goals. She is also friendly and dedicated to provide the students with the comfortable and pleasant working environment. I could not have imagined a better mentor.

I would also like to thank my committee members, Dr. Miroslaw Truszczynski , Dr. Chi Wang, Dr. Ruigang Yang and Dr. James N. MacLeod for their advisory and insightful suggestions. Their critical comments also make me to face the weakness of the methods and lead to a necessary improvement.

Special thanks to Dr. Jan F. Prins who is a collaborator and also the advisor during my visit in the computer science department in UNC-Chapel Hill. We have collaborated on several projects. His solid backgrounds in algorithms and computing help me finalize the methodology of my thesis. He also taught me how to work with researchers from diverse backgrounds. I really appreciate the experience working with him.

I thank my friends in University of Kentucky. It was truly a great pleasure I spent these years with you. Thanks you for letting me share my problems and excitements. I remember all the laughs and joys and the days we help each other. I am happy our

friendship extends beyond the days in UKY. Finally, I thank my family: my parents for their unconditional love and faith in me; my wife who has been with me for five years for her support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

This year 2013 is the ten year anniversary of the completion of the Human Genome project. The exploration on human genome in the past decade has revealed much about its sequence and structure, as well as identification and annotation of many functional sequences, known as genes. These efforts on genome studies keep refreshing our vision into the mystery of the functioning and heredity of living organisms, which has been and remains one central mission of the life sciences. Moreover, the discoveries of the associations between genomic signatures and human diseases have further led to the proposal of revolutionizing concept of "genomic medicine" in healthcare.

More and more diseases are being defined at the molecular level, for their genetic causes and progression. In some tumor types, such as breast cancers and the leukemias, molecular biomarkers have exhibited promise for clinical decision making in addition to histologic classification. These biomarkers may help to distinguish the disease into prognosis subcategories, to predict the response of a patient to a particular therapy and to potentially inform treatment strategies. By pinpointing the characteristics of disease, precise drug usage may be enabled.

Despite the early achievements using genomic features such as gene expression[1], the performance of using molecular signatures to help disease prognosis has progressed slowly. This raises the urgent need of, beyond the sequence annotation of the genome, the functional investigations into the genome, which aim to unveil the mechanism

[1]The intensity of a gene's sequence observed in a cell.

of how genome sequences function and to fill the gap from genetic information to phenotypes and diseases.

Modern sequencing technologies as well as bioinformatics approaches are rapidly evolving and are providing unprecedented opportunities for such missions. In this chapter, we will articulate the specific biological question and computational challenges that we aim to address, and we will show the landscape of the work completed in this dissertation.

1.1 Differential analysis of mRNA transcriptome

The genes have been considered as the basic units of the hereditary material passed from parents to offsprings. They contain the genetic information that determines many characteristics of an organism, such as the eye color of a human being.

In eukaryotic organisms, such as humans, animals and plants, the genes do not directly run the cells, but through an intermediate step called *transcription*. The product of transcription is a set of *transcripts*. In transcription, one gene can be viewed as a combination of multiple functional parts, and through reassembling these parts the gene may generate multiple transcripts. Each transcript will then be responsible for a particular function in the cell. Beyond the presence of a transcript, the abundance of the transcript correlates with the strength of the controlling signal. Proper abundance of the transcripts balances the normal function of the cell. The *mRNA transcriptome*, the totality of the diversity and the abundance of the transcripts, then characterizes a cell at a particular time or under a particular condition.

The transcriptome is known to vary in response to cellular differentiation and

diseases. By comparing transcriptomes at different conditions, the difference in the abundance of transcripts may associate with the change of phenotype, hence may reveal the functional roles of these transcripts. This type of differential analysis between transcriptomes is called the *differential transcription analysis*.

1.2 The prominence of differential transcription analysis empowered by sequencing the mRNA transcriptome (RNA-seq)

The advent of massively parallel RNA sequencing (RNA-seq) has provided an unprecedented opportunity to comprehensively picture the entire transcriptome. RNA-seq directly samples and sequences from the mRNA transcriptome without dependence on predetermined sequence templates. This enables the discovery of novel transcripts not cataloged in existing knowledge and those specific to a group of individuals. At the mean time, the transcript-level sequencing has a high resolution spelling every nucleotide sampled and makes possible accurate quantification of the transcripts present in the transcriptome.

The application of RNA-seq in clinical usage has also been practical. Current RNA-seq technologies allow the profiling of a patient's transcriptome for a cost typically less than $1,000 in less than one week.

Utilizing transcriptome samples from RNA-seq, differential transcription analysis is now empowered the potential to reveal novel molecular biomarkers for human diseases, through comparison of transcriptomes from normal samples and diseased samples. For example, specific alternative transcripts and fusion transcripts[2] are

[2]A fusion transcript is an abnormal hybrid transcript formed by two transcripts from different

commonly found in the cancer transcriptomes [Maher et al., 2009, Berger et al., 2010]. Differential transcription analysis among tumor samples and healthy samples may reveal transcripts involved in tumor progression. Afterwards, therapies may be developed to block the growth and spread of cancer by targeting on the identified molecules, for example, hampering cell growth signaling, promoting the specific death of cancer cells, stimulating the immune system to destroy specific cancer cells, and delivering toxic drugs to cancer cells [National Cancer Institute].

1.3 The current RNA-seq-based differential transcription analysis

1.3.1 A standard workflow

The direct product of an RNA-seq experiment is a digital file containing the reads from the sample and their sequences. Recognized by its computational difficulty and extraordinary volume, this new type of biological data has raised many computational and methodological challenges that excite the field of computational biology and bioinformatics.

The most straightforward approach for differential transcription analysis is to first measure the abundance of the transcripts using the sampled RNA-seq reads and then compare the abundances across samples. However, the genomic location where a read is sampled is not known, and the short reads cannot directly identify the original transcripts.

Therefore, the computational solutions for differential transcription analysis aim to bridge short read sequences and differences between original transcriptomes.

genes concatenated together.

Step 1: alignment of RNA-seq reads to the genome.

In order to profile the original transcriptome, it is necessary to know the relation among the sampled reads, such as their relative positions. This can be done by mapping the sampled reads onto a reference genome and representing them using the genomic coordinates. A reference genome is a set of known DNA nucleotide sequences that specify all known genes and intergenic regions of a species. It is assembled by sequencing the DNA from some representative individuals. Despite individual modifications such as single-nucleotide polymorphisms (SNPs) and insertions/deletions (indels), the reference genome is highly consistent across individuals in the species. The mapping of the RNA-seq reads, essentially, tries to find matches on the reference genome for the sequences of the sampled short reads (detailed in Section 7.2). Importantly, gaps are allowed in the read alignments. A read whose sequence can directly match a piece of the reference genome is mapped as an entirety and has an *unspliced alignment*. A read whose sequence should be split into segments and matched to different places of the reference genome separately has an *spliced alignment*. The spliced RNA-seq read alignments then suggest the splice junctions and introns on the genome, including those not known in existing annotation database.

Step 2: reconstruction of the transcripts

Provided the RNA-seq read alignment, the *ab initio* transcript reconstruction approaches can then be applied to reconstruct the transcripts in the original transcriptome (detailed in Section 7.4.1). Because the sequences of RNA-seq reads are those

kept in the mRNA transcripts, the genomic coordinates which have RNA-seq reads aligned to can help recover the exonic sequences on the genome. Nucleotides may be considered forming an exon if they are contained in a same read or a same mate-pair, or if they locate close to each other on the genome (*e.g.*, several bases apart) with no spliced alignments in between. The splice junctions will indicate how the exons should be concatenated during splicing. Then a graph can usually be constructed to picture the connectivity in a potential gene of all the reconstructed exons or of all the read alignments. Transcripts may further be identified by traversing the graph for confident graph paths.

Step 3: transcript abundance estimation

From the genomic alignments of the RNA-seq reads, the expression level of each gene may be evaluated, by collecting the number of reads falling in the gene's region. However, the abundance of individual transcripts is not trivial because the original transcript of each RNA-seq read is unknown. The transcript abundance estimation then aims to infer the proportions of the transcripts in the original transcriptome. Essentially, the genomic locations of the read alignments are regarded as the observations. The transcript percentages are related to the observed reads through the likelihood of a read being sampled from each transcript. This relation is particularly crucial for the reads aligned to exons shared by multiple transcripts. The result of this step will be the abundance of all transcripts reconstructed, solved by maximizing the joint probability of all observed reads.

Step 4: differential transcription analysis

With the set of transcripts reconstructed and their abundance estimated, the difference between two transcriptomes may be derived by explicitly comparing the diversities and expression profiles of the transcript sets. The abundance of every transcript is normalized across sample groups and statistically tested for equivalence, resulting in a list of genes with significant change of transcription from one condition to another. This strategy may provide direct insights into differentiated transcripts, but the accuracy of abundance estimation is often concerned.

1.3.2 The limitations of existing differential transcription analyses

Despite the large variety of computational methodologies developed and their success in RNA-seq-based differential transcription studies, existing approaches often suffer from unsatisfied robustness, limited accuracy, inefficient performance and poor reproducibility. These issues partially result from methodological shortages and implementation weaknesses that are specific to individual approaches. The major concern, however, is the deviation between the theoretical assumptions/requirements of the approaches and the limitations by the short read sequencing technologies.

Fundamentally, the sequencing capabilities of current RNA-seq protocols still limit the direct profiling of transcriptome using raw RNA-seq reads. Ideally, the sequencing procedure should reveal the diversity and abundance of a transcriptome in a complete, accurate and unbiased manner. However, the length of current RNA-seq reads is insufficient. The transcripts in human transcriptome have a median length around

2,500bp, a length much longer than an RNA-seq read (typically shorter than 100bp for a single-end read or about 250bp for a paired-end read). Because transcripts in a same gene may share a large amount of exonic sequences, it is highly ambiguous to identify the original transcript for reads sampled from the shared exons. Hence both the identity and the abundance of the transcripts may not be directly derived.

For example, current transcript level differential transcription analyses rely on accurate transcript reconstruction and abundance estimation. Both tasks, nonetheless, are known to be inaccurate and unstable, because current read length is not sufficient to identify the original transcripts with controlled ambiguity. The downstream differential analysis may perform poorly as a result.

In order to improve sensitivity and specificity, some differential analysis approaches on alternative splicing events have been proposed to perform local estimation and testing based on known alternative splicing patterns extracted from reference transcriptome. But this strategy sacrifices the power to detect novel splicing isoforms and often the generality to uncataloged or complex splicing patterns.

Moreover, the sequence file of every single sample may take up to tens of Gigabytes even in its binary format, consisting of hundreds of millions of read records. At the mean time, nowadays clinical RNA-seq datasets may grow larger and larger. For example, the Cancer Genome Atlas (TCGA) breast cancer analyzing project has sequenced more than 1,000 RNA-seq samples. This further requires a superb efficiency in algorithm design and practical performance in addition to a high accuracy for any analyzing approach. Existing approaches, however, typically work with one-to-one comparison or datasets with small sample sizes ($e.g.$, < 10).

Envisioning the need of accuracy and scalability by large-scale RNA-seq studies, this dissertation has placed the focus on the reference genome-based *ab initio* workflow for differential transcription analysis, for the purpose of tracing from the massive RNA-seq reads back to the differences among original transcriptomes.

1.4 The topic of this dissertation

The goal of this dissertation is to develop a novel computational framework for precise transcriptome analysis on large-scale RNA-seq datasets. Starting with only the RNA-seq read alignments on the reference genome, the framework aims to accurately reconstruct the transcription models in the original samples, including both transcripts cataloged in existing knowledge and novel transcripts, and to highlight the transcription events that are differential across samples/groups. The developed methodologies are expected to alleviate the computational infeasibility that impairs existing transcript-based approaches, which leads to weaknesses in sensitivity, specificity and efficiency. Envisioned the advent of large, complex clinical RNA-seq datasets, the work will also explore the solutions that leverage information from hundreds or thousands of samples, improve computational scalability, and resolve complex transcription signals due to noise and errors in sampling and alignment. The application of the framework developed in this dissertation may provide insights into patterns of transcription regulations associated with cell development and diseases.

1.5 Contributions of this dissertation

We have developed a novel *ab initio* framework for transcriptome analysis on large-scale RNA-seq datasets, working from raw RNA-seq reads and requiring only a reference genome. Leveraging biological interpretability and computational feasibility, this framework provides accurate and robust detection of differential transcription at the level of alternative splicing events, events that capture differences between transcript isoforms. In this way, this approach circumvents the computational infeasibility of full-length transcript reconstruction and estimation. On the other hand, this data-driven approach relies on neither transcript annotation nor pre-determined alternative splicing template, and hence frees the limitation on known events, known category and low complexity that all restrict existing splicing event-based methods.

The completion of this dissertation may contribute to the transcriptomics research in the following aspects.

- A probabilistic scheme has been developed for the accurate alignment of paired-end RNA-seq reads. Unlike typical aligners which determine the most probable alignments relying on presumed co-location of mate-pairs, the developed scheme reconstructs the actual unsequenced segments between the end reads, based on the splice structure derived from all the sampled reads. The full transcript fragment alignments are inferred maximizing the joint probability of all reads and splice junctions. In addition to resolving ambiguous short read mapping, these fragment alignments enable higher coverage on the transcriptome and more accurate splice junction detection. This alignment algorithm can be further

extended to fusion junction validation, demonstrated by applications in breast cancer RNA-seq datasets.

- A precise and efficient *ab initio* pipeline has been developed for alternative splicing-level transcriptome profiling and differential transcription analysis, relying on only RNA-seq read alignments. This pipeline includes a suite of novel algorithmic and statistical methods, featuring a graph theory-based algorithm for alternative splicing events discovery, an inference procedure for splicing isoform abundance estimation and a non-parametric significance test for evaluation of differential transcription between/among groups. The transcriptome representation using alternative splicing modules, in particular, provides a unique strategy in genome-wide transcription analysis distinguished beyond the existing methodologies. Through extensive evaluation on both simulated and real-world datasets, this pipeline has demonstrated superior sensitivity and specificity utilizing current short-read RNA-seq technologies.

- Classical differential transcription tests are based on the assumption that individuals in each sample group have the same distribution of transcript profile, which may not hold in clinical data due to biological heterogeneity among patients. In complement to the group-wise differential analysis, we have also explored the application of data mining techniques in gene transcription pattern discovery without relying on predefined sample groups. A non-parametric clustering scheme has been developed on the basis of sample-specific transcription profiles, accompanied with a novel statistical scoring criteria for informative

cluster selection. Applied on clinical datasets when within-group biological variation is significant and when existing grouping may not well characterize the individuals, this clustering scheme has demonstrated the ability to automatically find subgroups of samples that exhibit consistent transcription patterns. This provides a unique solution that may help highlight biomarkers differentiating disease subtypes and help refine existing subtype definition.

- We have identified differential splicing patterns potentially associated with breast cancer subtyping, though successful application on the Cancer Genome Atlas (TCGA) breast cancer dataset. Leveraging more than 800 RNA-seq samples, this framework has demonstrated the efficacy of joint transcription analysis on large-scale clinical RNA-seq datasets. Novel solutions have also been provided to the challenges such as computational scalability, data representation and transcriptome reconstruction from complex, noisy read alignments.

- The open-source software packages for the algorithms developed in this framework are released and actively maintained, publicly available to the research community.

Copyright © Yin Hu, 2013.

Chapter 2 Biological background

In this chapter, we will review the biological background and the technology platform of the differential transcription analyzing methodologies.

2.1 The genome in living organisms

The hereditary information of a living organism is stored in the molecule named deoxyribonucleic acid (DNA). This hereditary information contains the complete set of genetic codes instructing the development and functioning of the organism. In human cells, the DNA molecules are known for their double helix structure (Figure 2.1). The two strands are two long biopolymers running in opposite directions to each other, with the direction specified by the 5' end and the 3' end. Each strand is constituted by units named *nucleotides*. There are four types of nucleotides distinguished by their component nucleobase, denoted as G, A, T and C (guanine, adenine, thymine and cytosine), respectively. In this anti-parallel structure, the two strands are further bound together by pairing of corresponding nucleobases – A bonds only to T and C bonds only to G. The pair of two nucleotides binding together across the two strands is called a *base pair* (bp). The DNA sequence in human genome, for example, consists of more than 3×10^9 base pairs. The DNA sequence of nucleotides is the actual code that keeps the genetic information and is duplicated through complementary read pairing.

In cells of eukaryotic organisms, the most of the DNA is located inside the cell

nucleus, organized into structures called *chromosomes*. For human, for example, there are 23 pairs of chromosomes, each consisting of 50 million to 250 million base pairs. The functional regions in DNA sequences are further divided into hereditary units named *genes*. Every gene corresponds to a specific location on a chromosome and is typically responsible for a particular characteristic in the organism. The are approximately 20,000 genes for human. The DNA sequence of a gene contains the regulatory sequences that control the expression of the gene, as well as sequences that hold the information for cell differentiation, cell functioning and heredity. The term *genome* is then used to refer to the complete set of genetic information, including sequences of genes that contain genetic code and other functional but non-coding sequences.

2.2 The mRNA transcriptome

Within living organisms, proteins are the molecules that perform actual functions, such as catalyzing metabolic reactions, responding to stimuli and transporting molecules. Different proteins differ primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes. A large variety of protein has been discovered in the human body, estimated around 100,000. This size is more than five times the number of genes that code for the proteins. The great difference between the diversity of genes and that of proteins is due to a process named *transcription*.

The genetic code in the DNA sequences is not directly used for protein synthesis, but first through the transcription from DNA to ribonucleic acid (RNA). An enzyme called RNA polymerase first recognizes and binds a promoter region of the gene,

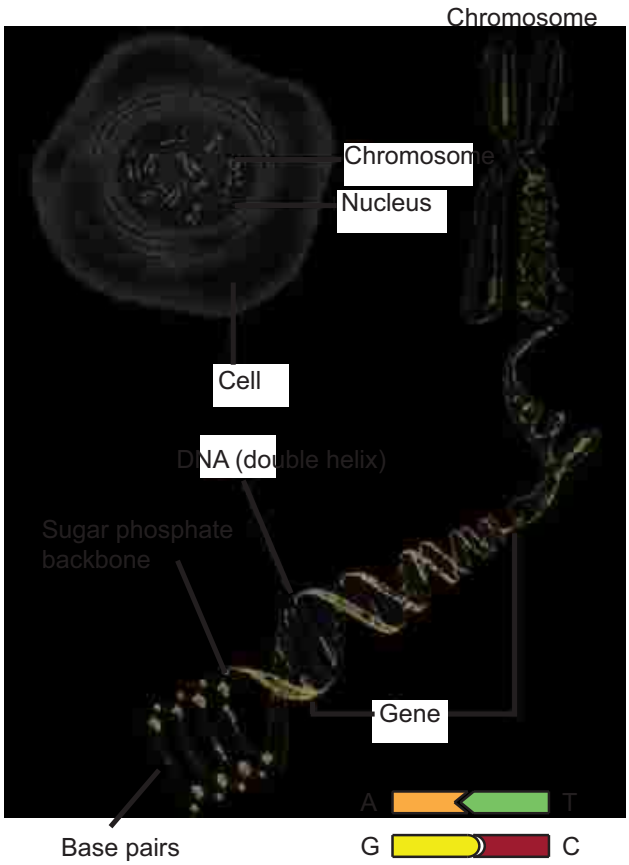


Figure 2.1: The genome in an eukaryotic cell. In eukaryotic organisms, such as animals, plants, fungi and protists, most of their DNA is stored inside the cell nucleus and is typically organized in chromosomes. Every gene is a piece of sequence on a chromosome, typically responsible for a particular characteristic in the organism. (Figure partially adapted from Wikipedia [Wik].)

unchains the double helix DNA structure, reads one specific strand from 3' to 5' and synthesizes the RNA from 5' to 3'. The synthesized RNA then has a nucleotide sequence matching the DNA sequence of the genome (the strand from 5' to 3'), except for that T is replaced by U. This single-stranded RNA molecule is known as the precursor mRNA (or pre-mRNA, primary transcript)

Not all nucleotides in the sequence of a pre-mRNA will constitute the genetic codes for protein synthesis. Sequences that are not responsible for the specification of amino

acids should be removed and not remain present within the final mature mRNA molecules. This post-transcriptional modification for the pre-mRNA molecules is called *splicing*. The removed sequences are named *introns*. The rest sequences are called *exons* and will be joined together (in transcription order) into the final mature mRNA after RNA splicing. Each mRNA molecule is also called a *transcript* or *isoform*. The procedure is shown in Figure 2.2.

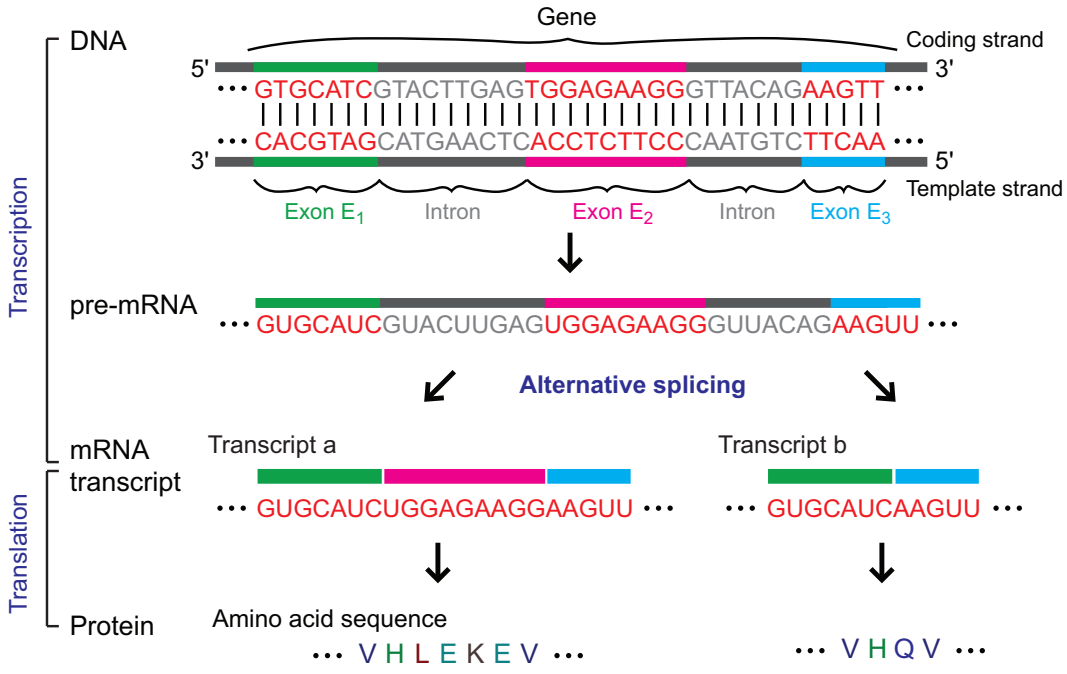


Figure 2.2: The alternative splicing and transcription. The DNA sequence of a gene is first copied to pre-mRNA, with all T's replaced by U's. The pre-mRNA is then spliced, removing introns and concatenating exons, to make mRNA transcript. One gene may code for multiple mRNA transcripts through alternative splicing. Different sets of exons, for example, $E_1E_2E_3$ and E_1E_2, may be retained to form different transcripts, transcripts a and b. Alternative transcripts will typically lead to different amino acid sequences. As a result, the produced proteins will have different amino acid compositions and structures, hence varied functions.

In many genes, the way to splice the pre-mRNA is not unique. The exons to be retained in the final mRNA are determined by the splice sites, regulated and

selected by trans-acting splicing activator and splicing repressor proteins. The intron regions are typically defined by consensus sequences in eukaryotic cells. An RNA and protein complex, known as the spliceosome, then binds to the specific consensus sequences [Clark, 2005]. The binding sites will then determine the ends of the intron to be spliced out and define the ends of the exon to be retained [Matlin et al., 2005]. Thereafter, the spliceosome will cleave the 5' end of the intron from the upstream exon and cleave the 3' end of the intron from the downstream exon. The two exons are joined and the intron is then released. [Lopez, 1998, Black, 2003] This process in which particular exons of a gene may be retained in or removed from the final mRNA transcripts of this gene is called *alternative splicing* (Figure 2.2). Through alternative splicing, the exonic sequences of one gene may be reconbined into different mRNA transcripts, hence one gene may code for multiple proteins. Due to different amino acid sequences, these proteins typically have different functions. In human genome, alternative splicing happens in more than 95% of multi-exon genes [Sultan et al., 2008, Wang et al., 2008, Pan et al., 2008, Kwan et al., 2008].

There are seven commonly observed forms of alternative splicing events (Figure 2.3). The most basic case is exon skipping, also known as cassette exon, in which one or multiple exons are alternatively included or excluded in the transcript. In mutual exclusive exons, each alternative exon is retained in one transcript after splicing. But no transcript may have both of the exons. For some consecutive exons connected by a splice junction, there may be an alternative sequence defining different boundaries of the 5' exon (donor side) or the 3' exon (acceptor side). Alternative splice junctions will then exist from alternative splice sites. In intron retention, a sequence

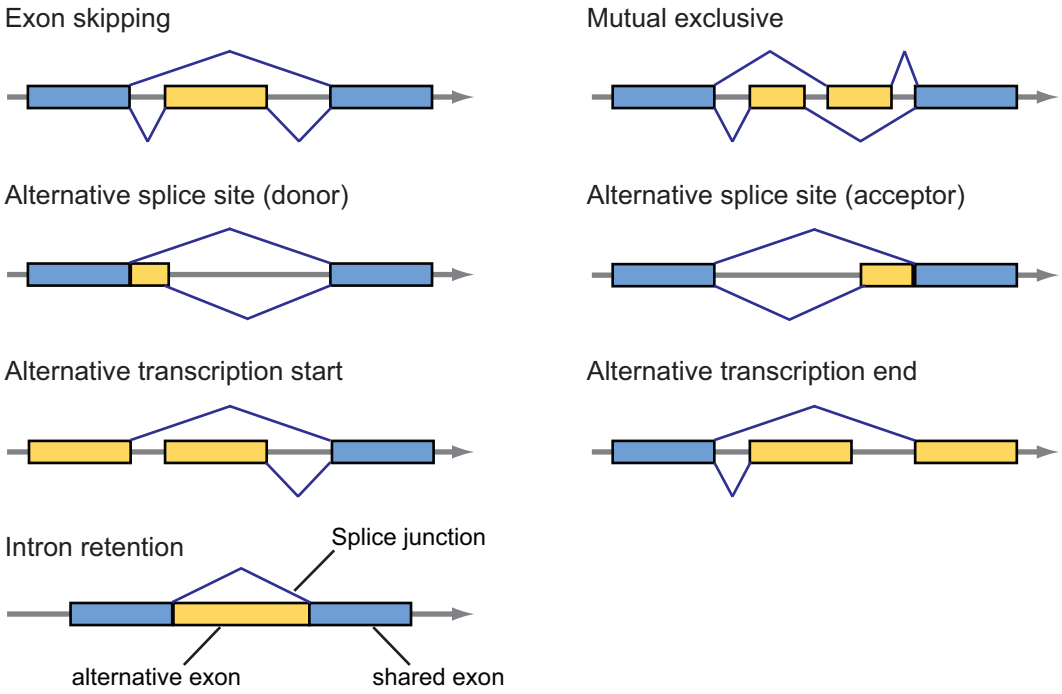


Figure 2.3: The common forms of alternative splicing events. The splice junctions indicate how the exons in the gene may be concatenated into different transcripts. The resulted transcripts share same blue exons but are distinguished by alternative exons colored yellow.

may either be retained in the mRNA transcript as an exon or spliced out as an intron. Furthermore, transcripts may also start or end at different exons, initiating or ending the transcription at different sites. More complicated alternative splicing events may be observed, combining the basic forms or developing more complex exon inclusion/exclusion patterns.

The totality of all mRNA transcripts transcribed from the genome within a functioning cell is referred to as the mRNA *transcriptome*, characterized by the diversity of the transcripts present and their individual quantities.

After transcription, the produced mRNA transcripts then serve as the templates for protein synthesis, in a process called *translation*. The nucleotide sequence of the

mRNA transcripts is decoded, linking specific amino acids into a chain. This amino acid chain will later fold into an active protein that carries functions in the cell.

2.3 Differential transcription

As the union of all protein synthesis templates, the mRNA transcriptome is often regarded as a precursor for the entire set of proteins that may be produced from the genome. The diversity of the mRNA transcriptome correlates with the diversity of the proteins. More importantly, the abundance of the mRNA transcriptome, the quantity of each individual transcript in the cell, may further associate with the expression levels of the proteins thus may have direct impact on the cell's function.

The diversity and abundance of transcripts transcribed from the genome are basic characteristics of a cell at a particular time under a specific condition, and are known to vary in response to cellular differentiation and maturation as well as environmental factors and disease. Comparing the transcriptomes may highlight the genes and transcripts that are being actively expressed, or vice versa. For example, transcripts whose corresponding proteins have functions related to cell proliferation may have shifted expression at different stages of cell development. Therefore, the difference between mRNA transcriptomes sampled at these stages may provide insight into the functional effects of cell differentiation and cell life cycles [Wang et al., 2008, Trapnell et al., 2010].

On the other hand, because the regulation of transcription controls the intensity of particular proteins, abnormalities in alternative splicing and the resulted transcripts often lead to diseases [Faustino and Cooper, 2003, Tazia et al., 2009]. For example,

associations have been reported between irregular alternative splicing and heart diseases [Xu et al., 2005], muscle diseases [Poulos et al., 2011], Parkinson disease [Fu et al., 2013], neurological diseases [Yap and Makeyev, 2013, Poulos et al., 2011], *etc.* In addition, the alternative splicing in cancer-related genes often has important roles in cell proliferation and tumor suppression, further associated with various types of cancer such as breast cancer [Tammaro et al., 2012], ovarian cancer [Wang et al., 2010b], lung cancer [Pio and Montuenga, 2009] and prostate cancer [Haile and Sadar, 2011]. Therefore, the difference between mRNA transcriptomes sampled from healthy and diseased cells may provide insight into the functional consequences of disease, as well as help to identify biomarkers that can classify different disease types [Wang and Cooper, 2007].

The topic of this dissertation then focuses on the *differential transcription* analysis, the detection of differences between the transcriptomes at given times or conditions.

2.4 Traditional approaches for transcriptome study

Differential gene expression analysis is the first attempt to correlate the function of cells with genes active or inactive. The quantitation of the genes' expression levels may further allow downstream pathway analyses to seek target genes for diseases. DNA Microarray technology [Clark et al., 2002, Russo et al., 2003] has been used as a powerful tool to quantitatively measure the expression levels of thousands of genes simultaneously, enabling the comparison of gene expression among multiple samples.

The microarray technologies rely on the hybridization between two DNA strands,

the specific pairing of complementary nucleic acid sequences. A large collection of microscopic DNA probes is attached to a solid surface, called an array. Each probe (feature) encodes a specific DNA sequence that may identify a gene or other DNA element, and the identify of the probes is known by their position. A nucleic acid sample (target) is fluorescently labeled, hybridized to the probe sequences, and washed after hybridization. A higher number of complementary base pairs makes tighter bonding between two nucleotide sequences. The non-specific bonding sequences may be washed off. The target sequences that bind will generate a signal. The relative expression of a genes in the sample may then be quantified by measuring the intensity of the signal at corresponding probe, which correlates with the amount of target sample binding to that probe. [National Center for Biotechnology Information, National Human Genome Research Institute]

Transcript-level differential expression has been receiving more and more interests, at a higher resolution, than the differential gene expression analyses. The microarray technologies were then applied on exons, called exon arrays [Okoniewski and Miller, 2008, Xi et al., 2008], to detect differences in the expression of known gene exons, which may reveal expression of individual transcripts that retains or skips the exon.

In addition to microarrays, molecular techniques like CAGE (Cap analysis gene expression) [Shiraki et al., 2003] and SAGE (Serial analysis of gene expression) [Velculescu et al., 1995] have also been developed to determine the transcription start position and transcript expression, provided the transcript sequence. Small fragments of transcripts are extracted and sequenced to generate a set of short nucleotide sequences (tags) that may identify their original transcripts. With the guidance of

a reference genome the expression of the transcripts may be measured through the observed counts of the tags.

The common limitation of these traditional technologies is the requirement of pre-known sequence, allowing analyses only on annotated genes/exons/transcripts. The microarray technology is also affected by the background noises [Tu et al., 2002, Klebanov and Yakovlev, 2007] and the limited dynamic range of expression levels (typically up to 10^2) [Wang et al., 2009], degrading the accuracy of the measurement.

2.5 The revolutionizing RNA-seq technology

More recently, high-throughput sequencing methods such as RNA-seq [Wang et al., 2009, Pan et al., 2008, Wang et al., 2009] have been able to accurately record short sequences of nucleotides sampled from millions of mRNA molecules in the transcriptome, and thereby are capable of observing samples from known as well as unknown transcripts. Through massively sequencing the whole transcriptome, RNA-seq has made possible an accurate and comprehensive snapshot of the mRNA transcriptome. For example, Illumina's Hiseq2000 can produce up to 200 million reads in one lane in one sequencer run. The large number of molecules randomly sampled provide the potential to not only characterize the diversity of transcripts present in the transcriptome but also accurately estimate the relative abundance of transcript isoforms.

In a typical RNA-seq experiment, the mRNA molecules in the target transcriptome will be first synthesized into cDNA, followed by a process of random fragmentation that cuts the full-length mRNA transcripts into shorter fragments. A size selection process will then be performed, and fragments with a proper size (typi-

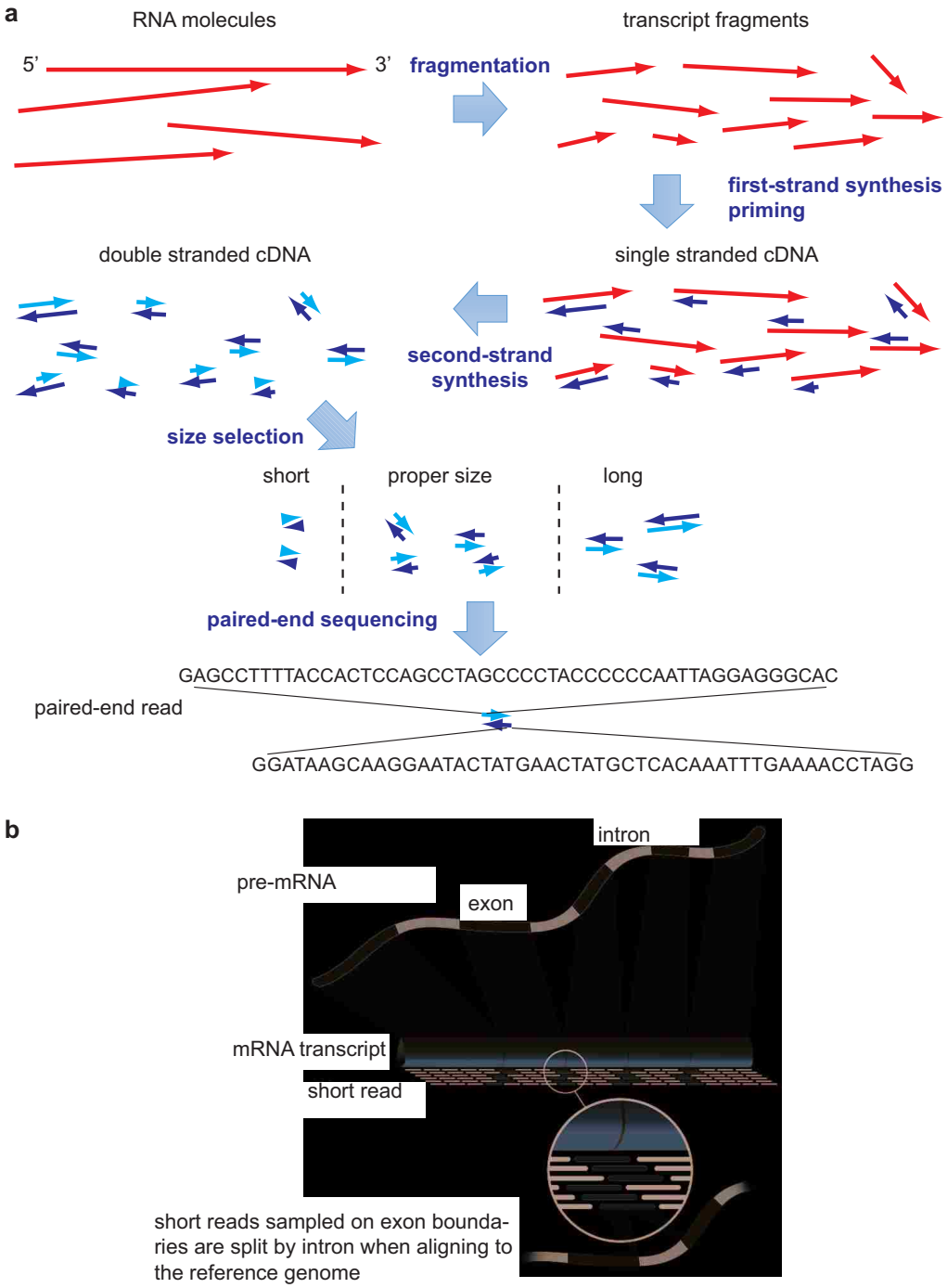


Figure 2.4: An illustration of the high-throughput mRNA sequencing. **(a)** The steps of a typical RNA-seq experiment. The RNA molecules will be synthesized into cDNA, fragmented and size-selected, before getting sequenced from one end or both ends. **(b)** The discovery of exons and exon-exon splice junctions using RNA-seq short reads. Directly sampling on the mRNA transcripts, RNA-seq needs no guidance from pre-known sequences and may reveal splice junctions as well as transcript isoforms cataloged or novel, well-abundance or lowly-expressed. (Figure partially adapted from Wikipedia [Wik].)

cally from 300bp to 500bp) will be selected for sequencing. The direct output of an RNA-seq experiment, afterward, will be tens or hundreds of millions of short reads, typically of length less than 100nt, whose sequences are read from the transcript fragments in the prepared cDNA library. In order to improve sampling coverage at the same sequencing capability, the paired-end sequencing strategy has been widely adopted, which sequences both ends of a transcript fragment. The additional pairing information can help determine the entirety of the fragment, benefiting the analyses of transcriptome through the guidance of, for example, read alignment and original transcript identification. (Figure 2.4a)

Unlike traditional approaches which may only work with known genes or known transcripts, RNA-seq does not rely on any pre-specified sequences or pre-determined templates. It makes possible the discovery of previously uncataloged exons and splice junctions and hence novel transcripts and genes (Figure 2.4b). Compared to microarrays whose detection is limited by the scope of the probes, the transcript fragments sequenced by RNA-seq, in principal, are randomly sampled, which allows an unbiased and comprehensive survey of the transcriptome. Furthermore, the broad dynamic range of RNA-seq enables the analysis not only on abundant transcripts but also transcripts that are barely expressed, providing a more accurate expression measurement than microarrays.

Copyright © Yin Hu, 2013.

Chapter 3 A Probabilistic Model for Aligning Paired-end RNA-seq Data

3.1 Introduction

High-throughput sequencing technologies are providing unprecedented visibility into the mRNA transcriptome of a cell. In cancer, alternative splicing and gene fusion events [Maher et al., 2009, Berger et al., 2010] are common changes observed in the mRNA transcriptome. Cancer specific splicing events are promising biomarkers and targets for diagnosis, prognosis, and treatment purposes. Recently, several computational methods [Trapnell et al., 2009b, Au et al., 2010] have been developed to identify splicing events using RNA-seq data. These methods align RNA-seq reads to the reference genome rather than to a transcript database, making it possible to identify novel splicing events via gapped alignment of reads to the genome.

New protocols and sequencing methods have expanded the length and type of RNA-seq reads, enabling more accurate characterization of the splices present in the transcriptome. A *single read* may constitute 35 – 100 consecutive nucleotides of a fragment of an mRNA transcript. The *paired-end read* (PER) protocol sequences two ends of a size-selected fragment of an mRNA transcript and reports the results as a pair. The fragment length is typically around 200bp but may vary according to different PER protocols. In our experiment, for example, the expected size of mRNA fragments are around 182bp ($\pm$40bp). Both ends of the fragment are sequenced to

at least 35bp in length.

This chapter focuses on predicting the alignment of an entire PER fragment, starting from the alignments of its end reads and using the alignments of other overlapping PER end reads to predict an overall alignment consistent with the expected length of the fragment. Since a PER fragment can be longer than single reads sequenced with today's RNA-seq technology, achieving such alignments may significantly increase the effective transcriptome coverage. Longer alignments also decrease alignment ambiguity in regions with genome repeats.

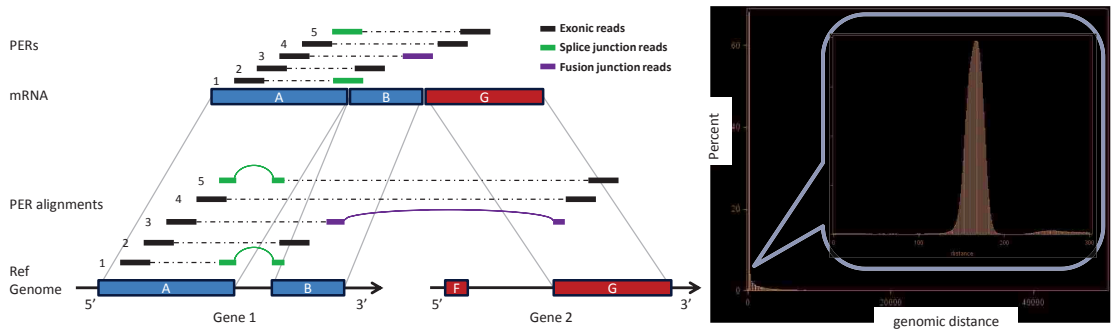


Figure 3.1: Left: A fragment of an mRNA transcript exhibiting gene fusion between exon B in gene 1 and exon G in gene 2 is sampled by six paired-end reads. The alignment of the transcript to the reference genome as well as the alignment of the PERs to the genome is shown. The unsequenced segments of PERs can not readily be aligned to the genome because of unknown intervening splicing events including, in this case, the fusion junction. Right: An example of the distribution of distance in genomic coordinates between paired end-read alignments generated from 2x35bp PER data. While the majority of distances fall within the normal distribution for mate-pair distance on mRNA fragments, a significant portion of the distances are far beyond the expected range, indicating potential splicing events.

A unique challenge in *PER fragment alignment* is that the expected distance between the two end reads within the trancript fragment, known as *mate-pair distance*, can be very different from distance between the two end reads when aligned to the genome. This can happen when the two ends fall in different exons, so that their sep-

aration in genomic coordinates includes one or more intervening introns that are not present in the transcript (Figure 3.1 left). This effect is illustrated as a long tail in the mate-pair distance distribution when aligned on the genome (Figure 3.1 right). Resolving the discrepancy between the expected mate-pair distance and the paired-end separation on the genome is not trival. RNA-seq aligners including TopHat [Trapnell et al., 2009b] and SpliceMap [Au et al., 2010] align PERs using heuristics. When the distance between end alignments is substantially longer than the expected mate-pair distance, TopHat reports the closest end alignment for a PER, while SpliceMap considers PERs with ends mapped within 400,000bp on the genome. While both heuristics have meaningful biological motivations, neither method predicts or validates the PER alignment. Since both approaches discard PER alignments that span a very long interval or cross chromosomes, neither of them is capable of finding long range splicing or gene fusion events.

In this chapter, we have proposed a new probabilistic framework for aligning RNA-seq PERs to a reference genome, without relying on transcript databases. Our goal is to discover both short range splice junctions and long range splice/fusion junctions through accurate mapping of PER end reads as well as the unsequenced middle portion. Our approach starts by building a compact splice graph to represent all putative splicing events, regardless of the intron sizes, derived from individual end read alignments. An expectation-maximization algorithm is then applied to identify the most probable path in the graph that connects the two ends of a PER based on the empirical distribution of the mate-pair distances. This in turn is used to infer the significant splice junctions.

Our approach was applied to RNA-seq data sets of 2x35bp PER reads from MCF-7 and SUM-102, two well known breast cancer cell lines. PER fragment alignment increased the coverage three fold compared to the alignment of the end reads alone, and increased the accuracy of splice detection. The accuracy of the EM algorithm in the presence of alternative paths in the splice graph was validated by qRT-PCR experiments on 8 exon skipping alternative splicing events. PER fragment alignment with long range splicing confirmed 8 out of 10 fusion events identified in the MCF-7 cell line in an earlier study by [Maher et al., 2009].

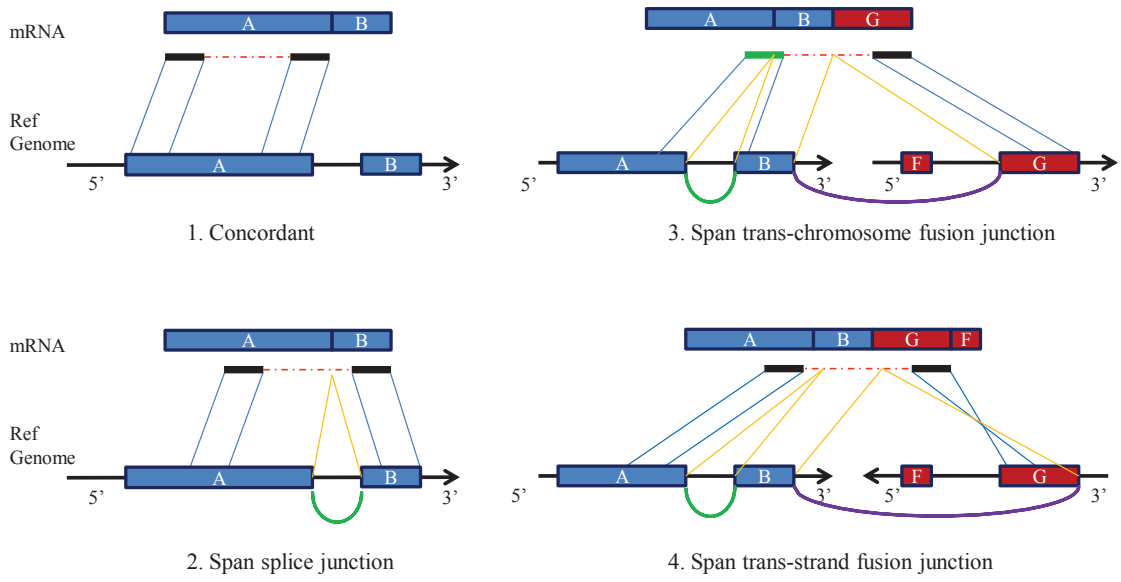


Figure 3.2: An illustration of a PER fragment alignment to the reference genome. The mRNA transcript is shown at the top, the paired-end read sequence is shown in the middle, and the alignment of the paired-end read to the genome is shown at the bottom. Four cases are shown: (1) concordant with mRNA alignment distance, (2) crossing a splice junction, (3) crossing trans-chromosome fusion junction, (4) crossing trans-strand chimeric junction.

3.2 Mapping Individual Reads

The alignment of RNA-seq PERs starts with the alignment of their individual end reads. MapSplice [Wang et al., 2010a] was used to map these end reads to the reference genome, generating both the read alignment and putative splice and fusion junctions.

MapSplice finds both exonic and spliced alignments of RNA-seq reads to a reference genome without any dependence on annotations or structural features of the genome. MapSplice operates by partitioning RNA-seq reads into short segments (18 – 25bp) that are aligned directly to the reference genome. Segments that can be aligned in this fashion are likely to be transcribed from exonic regions. Segments that cannot be aligned in the first step may contain a splice junction that is located by a search extending from aligned neighboring segment(s). In general, each segment may end up with multiple alignments that exceed some alignment quality threshold σ. A merge phase constructs candidate alignments for each read, by combining consistent alignments of its segments. Splice junctions are given a confidence value by considering the quality and diversity of all candidate alignments that include the junction. Finally the candidate alignments for a read are restricted to those in which the overall alignment quality and the confidence of any included splice junctions exceeds σ. The alignment of end reads by MapSplice was performed globally, i.e. without constraints on the proximity or strand of the mate-pair alignments in genomic coordinates. Given a PER (x_α, x_β), the alignments of x_α and x_β fall into one of the following four categories.

1. x_α and x_β are mapped onto the same chromosome and the same strand, and the mapped distance on the genome is close to their expected mate-pair distance (as shown in Figure 3.2(1)).

2. x_α and x_β are mapped onto the same chromosome and the same strand with a distance much longer than the expected mate-pair distance. This indicates the x_α and x_β span distinct exons (as shown in Figure 3.2(2)). When the distance is larger than 50,000bp, the two reads are assumed to be from different genes. Similar rules were used by Maher et al. [2009].

3. x_α and x_β are mapped onto different chromosomes. This indicates a potential trans-chromosome fusion event (as shown in Figure 3.2(3)).

4. x_α and x_β are mapped onto different strands, either of a same chromosome or of different chromosomes. This indicates a potential trans-strand chimeric event (as shown in Figure 3.2(4)).

In the first category, the alignment of a PER fragment can easily be determined since their separation is concordant with the expected mate-pair distances. For the remaining categories, the alignment of the complete PER fragment requires knowledge of the intervening exons and splicing structure to reconstruct plausible alignments. The set of splice junctions can be inferred from the spliced alignment of PER end reads. Reads 1 and 3 in Figure 3.1 left are examples of splice junction reads, while read 5 is an example of a fusion junction read. However, due to alternative splicing, multiple splicing paths may exist from x_α to x_β. Furthermore, the mapping of in-

dividual end reads may have multiple alignments to the genome due to repeats and homologous genes. To address these problems, we propose a maximum likelihood approach to disambiguate the PER alignments, detailed in the next section.

3.3 Probabilistic Framework

3.3.1 Graphical model and Notations

The spliced alignments of individual end reads result in a putative set of splice and fusion junctions. These junctions can be used to build a splice graph $G = \langle V, E \rangle$ to reflect the relation between the genome and transcript fragments. Within the splice graph G, each node $v \in V$ corresponds to a base on the reference genome. The nodes are connected by directed edges in the direction of the transcription. There exist two types of directed edges. The first type represents the connections between two adjacent bases on the same chromosome. The second type of edge corresponds to splice or fusion junctions, and skips around the spliced-out portion of the genome.

Let D be the set of RNA-seq PERs. Let x_α and x_β be the two end reads of transcript fragment x, $\langle x_\alpha, x_\beta \rangle \in D$. We denote the unsequenced segment of x as x_γ. Therefore, the entire PER fragment of x is the concatenation of x_α, x_γ and x_β and must be arranged in precisely this order, i.e., $x = \langle x_\alpha, x_\gamma, x_\beta \rangle$. Figure 3.3 illustrates the alignment of a PER based on the constructed splice graph. We are interested in predicting the alignment of entire fragment x including unsequenced x_γ as well as x_α and x_β.

Let Π_x^α and Π_x^β be the sets of valid alignments of end reads x_α and x_β, respectively.

The set of putative end read alignments of a PER contains all the unique combinations of the mapped locations of x_α and x_β, i.e.,

$$\Pi_x^{\alpha,\beta} = \{\langle \pi_x^\alpha, \pi_x^\beta \rangle | \pi_x^\alpha \in \Pi_x^\alpha, \pi_x^\beta \in \Pi_x^\beta\}.$$

Determining the alignment of x_γ is not straightforward since it is not sequenced. Its alignment might be predicted given the mapping of the end reads π_x^α and π_x^β and the splicing paths connecting them. We use $\Pi_x^{\gamma|\alpha,\beta}$ to denote the set of candidate alignments of x_γ given π_x^α and π_x^β, each of which corresponds to a unique concatenation of exonic regions by following a particular splicing path. A putative alignment of a PER x, π_x, therefore, is equivalent to an acyclic path that starts with the first base of π_x^α, passes π_x^γ and ends with the last base of π_x^β. Formally, given the set of end read alignments $\Pi_x^{\alpha,\beta}$, the set of candidate alignments of x, Π_x, is

$$\Pi_x = \{\pi_x | \pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}, \pi_x^\gamma \in \Pi_x^{\gamma|\alpha,\beta}\}.$$

Problem Definition: Let $\Pi = \{\pi_x | \langle x_\alpha, x_\beta \rangle \in D\}$ be the set of candidate fragment alignments for all PERs in D. Our goal is to determine an alignment for each PER, $\hat{\Pi}$, that maximizes the likelihood of the alignment of all the PERs in D, i.e.,

$$\hat{\Pi} = \arg\max_{\Pi} \prod_{x \in D} P(x|\Pi). \tag{3.1}$$

3.3.2 Probability definitions

Probability of a PER: The probability of a PER x is determined by its end read alignments $\Pi_x^{\alpha,\beta}$. By summing up the probability that a read alignment $\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}$

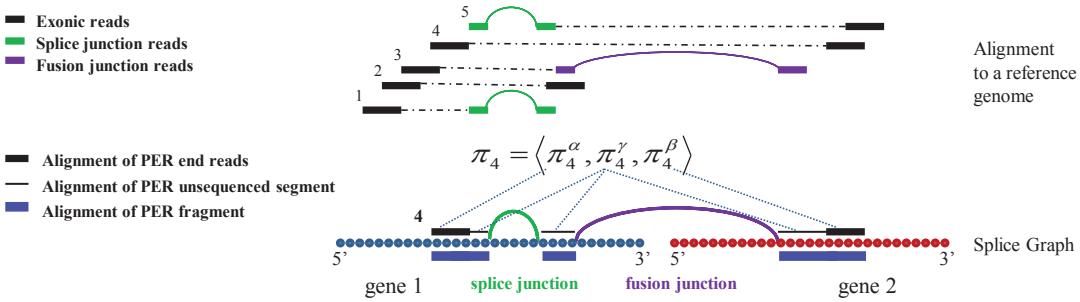


Figure 3.3: An illustration of the framework proposed in Section 3.3 applied to the example in Figure 3.1. The input is a set of RNA-seq PERs that have both ends aligned to the reference genome (top row). A splice graph can be constructed by taking each base as a node and connecting adjacent bases in the same chromosome as well as bases that constitute a potential splice junction or fusion junction (second row). A candidate alignment of a PER is a path in the splice graph from its start position to end position with the proper orientation.

is the true alignment $\hat{\pi}_x^{\alpha,\beta}$ at each candidate alignment in $\Pi_x^{\alpha,\beta}$, the probability of x can be computed as

$$\begin{aligned} P(x) &= \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \\ &= \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(x|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \cdot P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}). \end{aligned}$$

Here $P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ is the expected probability that x is aligned to $\pi_x^{\alpha,\beta}$. It is estimated at the expectation step of EM algorithm described in Section 3.3.3. The probability of x's alignment given $\pi_x^{\alpha,\beta}$, $P(x|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$, is determined by

- the probability of the accurate alignments for both end reads, x_α and x_β;
- the probability of the alignment for unsequenced portion, x_γ.

Mathematically, assuming the assessment of x_α, x_β and x_γ are independent,

$$\begin{aligned} P(x|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) &= P(x_\alpha|\hat{\pi}_x^{\alpha} = \pi_x^{\alpha}) \cdot P(x_\beta|\hat{\pi}_x^{\beta} = \pi_x^{\beta}) \\ &\quad \cdot P(x_\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}). \end{aligned}$$

We first determine $P(x_\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$, the probability of x_γ given $\pi_x^{\alpha,\beta}$. x_γ is the unsequenced portion of x. Its alignment, π_x^γ , would be one of the putative splicing paths connecting π_x^α and π_x^β, assuming the necessary splice junctions are present. Since the length of π_x^γ corresponds to the *mate-pair distance*, for each putative alignment π_x^γ, the probability $P(x_\gamma|\pi_x^{\alpha,\beta,\gamma})$ may be determined by the length of π_x^γ in the empirical distribution of the mate-pair distances $\mathcal{N}_d$. Here we denote it as $P_d(\pi_x^\gamma)$. Therefore, the probability of x_γ given end read alignment $\pi_x^{\alpha,\beta}$ can be expressed as

$$
\begin{aligned}
& P(x_\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \\
= & \sum_{\pi_x^\gamma \in \Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma, \hat{\pi}_x^\gamma = \pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \\
= & \sum_{\pi_x^\gamma \in \Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma|\hat{\pi}_x = \pi_x^{\alpha,\beta,\gamma}) \cdot P(\hat{\pi}_x^\gamma = \pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \\
= & \sum_{\pi_x^\gamma \in \Pi_x^{\gamma|\alpha,\beta}} P_d(\pi_x^\gamma) \cdot P(\hat{\pi}_x^\gamma = \pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})
\end{aligned}
$$

where $P(\hat{\pi}_x^\gamma = \pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ is the probability of an alignment π_x^γ given the end read alignment $\pi_x^{\alpha,\beta}$. We will determine $P(\hat{\pi}_x^\gamma = \pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ during the maximization step of EM algorithm described in Section 3.3.3.

The probability of the sequenced end reads, $P(x_\alpha|\hat{\pi}_x^\alpha = \pi_x^\alpha)$ and $P(x_\beta|\hat{\pi}_x^\beta = \pi_x^\beta)$ should be evaluated first based on their alignments, i.e., the probability of an alignment is not erroneous given their sequence similarity to the reference genome and their base call quality score (Li et al., 2008b), denoted as $P_q(x_\alpha|\pi_x^\alpha)$. In case a read spans one or more splice junctions or fusion junctions, the probability of a read is also dependent upon the joint probability of these junctions. Let $\Lambda(\pi_x^\alpha)$ be the set of junctions spanned by the end read alignment π_x^α. Considering both the spliced align-

ment and the matching quality, the probability of an accurate end read alignment can be calculated as

$$\begin{aligned} & P(x_\alpha|\hat{\pi}_x^\alpha = \pi_x^\alpha) \\ = & \begin{cases} P_q(x_\alpha|\pi_x^\alpha) & , \Lambda(\pi_x^\alpha) = \emptyset; \\ P_q(x_\alpha|\pi_x^\alpha) \cdot \prod_{\lambda \in \Lambda(\pi_x^\alpha)} P(\lambda) & , \text{otherwise.} \end{cases} \end{aligned}$$

The probability $P(x_\beta|\hat{\pi}_x^\beta = \pi_x^\beta)$ can be calculated similarly.

Splice Junction Probability: Splice junctions are derived from the spliced alignment of end reads to the reference genomes without relying on existing annotations. Such approach enables us to discover novel junctions but some of these junctions might be false positives. For example, if a junction has few and/or low probability PER supports, it may be spurious. On the other hand, a junction is likely to be true if it is crossed by at least one PER alignment with high probability. Therefore, we may evaluate the probability of a junction based on the set of PERs crossing it.

Mathematically, let $\Pi(\lambda)$ be the set of PER alignments going through the junction λ,

$$\Pi(\lambda) = \{\pi_x | \lambda \text{ is crossed by } \pi_x\}.$$

For each alignment $\pi_x = \pi_x^{\alpha,\gamma,\beta}$ in $\Pi(\lambda)$, the junction λ may be crossed in a spliced alignment of either π_x^α and π_x^β or be part of the splicing path of π_x^γ.

The *probability of the junction* λ can be expressed as the probability that there is

at least one PER alignment π_x in $\Pi(\lambda)$ supporting the junction, i.e.,

$$\begin{aligned} P(\lambda) &= 1 - \prod_{\pi_x \in \Pi(\lambda)} (1 - P(x, \hat{\pi}_x = \pi_x)) \\ &= 1 - \prod_{\pi_x \in \Pi(\lambda)} (1 - P(x|\hat{\pi}_x = \pi_x) \\ &\quad \cdot P(\hat{\pi}_x = \pi_x)) \end{aligned}$$

where

$$P(x|\hat{\pi}_x = \pi_x) = P(x_\alpha|\hat{\pi}_x^\alpha = \pi_x^\alpha)P(x_\beta|\hat{\pi}_x^\beta = \pi_x^\beta)P(x_\gamma|\hat{\pi}_x^{\gamma|\alpha,\beta} = \pi_x^{\gamma|\alpha,\beta}),$$

i.e., the probability that x is true at the alignment $\pi_x = \pi_x^{\alpha,\gamma,\beta}$.

In the next section, we will discuss an expectation maximization approach that determines the alignment for each PER maximizing the probability of all PERs as in Equation 3.1.

3.3.3 Probability Estimation

In this section, we apply the EM algorithm [Dempster et al., 1977, Wu and Jeff, 1983] to maximize the log likelihood of all the sampled PERs. The dependency relationships of all the variables are summarized in Figure 3.4.

Initialization

The probability of a PER is dependent upon the joint probability of the junctions within the span of the PER end read alignment. And the probability of a junction is calculated based on the probabilities of the PERs supporting the junction. In order to start the maximization, we initiate the probability of each junction as 1, and calculate the probability of each PER alignment.

At the alignment $\pi_x^{\alpha,\beta}$, the probability that the PER x takes π_x^{γ} as the unsequenced segment alignment is initiated with the expectation

$$\begin{aligned} & P(\hat{\pi}_x^{\gamma} = \pi_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) \\ = & \frac{P(x_\gamma, \hat{\pi}_x^{\gamma} = \pi_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{\sum_{\tilde{\pi}_x^{\gamma} \in \Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma, \hat{\pi}_x^{\gamma} = \tilde{\pi}_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}. \end{aligned}$$

Meanwhile the expected probability that $\pi_x^{\alpha,\beta}$ is true is estimated by

$$P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) = \frac{P(x | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{\sum_{\tilde{\pi}_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(x | \hat{\pi}_x^{\alpha,\beta} = \tilde{\pi}_x^{\alpha,\beta})}. \tag{3.2}$$

Then the probability $P(x)$ of every PER x and the probability $P(\lambda)$ of every junction λ can be computed based on the initial estimation.

Maximization and Expectation

The likelihood of the data is based on the probability $P(\hat{\pi}_x^{\gamma} = \pi_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$. We define the function $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$,

$$\begin{aligned} & Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})) \\ = & \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}. \end{aligned}$$

The EM algorithm performs maximization and expectation iteratively. At each iteration, hill climbing algorithm is applied to estimate $P(\hat{\pi}_x^{\gamma} = \pi_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ for every PER x such that $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$ is maximized. The proof that the maximization of $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$ will lead to the maximization of $l(D)$ is included in the supplemental materials. At the end of each iteration, the probability that PER x is mapped to alignment $\pi_x^{\alpha,\beta}$ is updated by taking the expectation, as

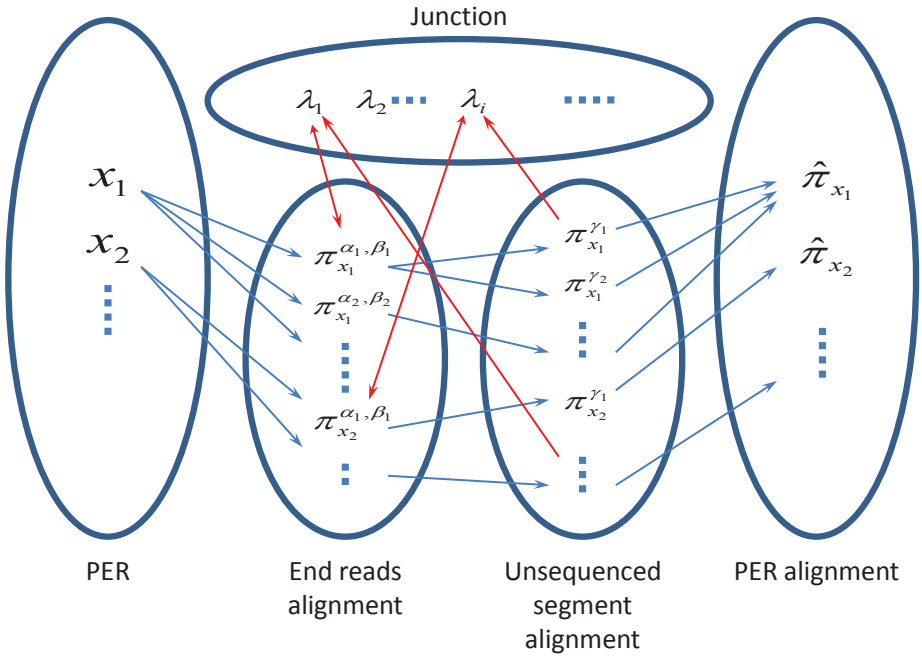


Figure 3.4: An illustration of the dependency relationship among the alignments of end reads, the alignments of unsequenced segments and junctions during the inference of the PER alignments. Within this probabilistic model, the probability of a junction is dependent on the PERs that support the junction, and the probability of a read alignment is dependent on the joint probability of the junctions spanned by the read alignment, as indicated with the red arrows. Taking PERs as input, our method aims at identifying the most probable alignments for every mate-pair x.

calculated in Equation 3.2. The proof of correctness for the EM approach is included in the supplemental materials.

Convergence of the EM algorithm

Abbreviate $P(\hat{\pi}_x^{\gamma} = \pi_x^{\gamma} | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ as θ. Let $q(\pi_x^{\alpha,\beta})$ be a "probability variable" about the event that x is mapped to alignment $\pi_x^{\alpha,\beta}$. The function Q is defined as

$$\begin{aligned} & Q(\theta, q, D) \\ = & \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}{q(\pi_x^{\alpha,\beta})}. \end{aligned}$$

We claim that the function Q has the following two properties.

Property 1.

$$l(\theta, D) \geq Q(\theta, q, D)$$

proof

$$\begin{aligned}
l(\theta, D) &= \sum_{x \in D} log \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta) \\
&= \sum_{x \in D} log \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}{q(\pi_x^{\alpha,\beta})} \\
&\geq \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}{q(\pi_x^{\alpha,\beta})} \\
&= Q(\theta, q, D)
\end{aligned}$$

Property 2.

$$l(\theta, D) = Q(\theta, q, D)|_{q = P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}$$

proof

$$\begin{aligned}
& l(\theta, D) - Q(\theta, q, D) \\
= & \sum_{x \in D} (log P(x|\theta) - \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}{q(\pi_x^{\alpha,\beta})} \\
= & \sum_{x \in D} (log P(x|\theta) \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) \\
& - \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}{q(\pi_x^{\alpha,\beta})} \\
= & \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) (log \frac{P(x|\theta) q(\pi_x^{\alpha,\beta})}{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}) \\
= & \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} q(\pi_x^{\alpha,\beta}) (log \frac{q(\pi_x^{\alpha,\beta})}{P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)}) \\
= & \ 0, \text{ iff } q(\pi_x^{\alpha,\beta}) = P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta)
\end{aligned}$$

Within each iteration, the evaluation function Q is calculated. Let θ^i denote the parameter estimated in i-th iteration. Let τ^i represent $P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta} | \theta^i)$, the

expected probability that x is aligned to $\pi_x^{\alpha,\beta}$ based on θ^i.

$$\begin{aligned} & l(\theta^{i-1}, D) \\ = \; & Q(\theta^{i-1}, P(\pi_x^{\alpha,\beta}|\theta^{i-1}), D) \\ = \; & Q(\theta^{i-1}, \tau^{i-1}, D) \\ \leq \; & Q(\theta^{i}, \tau^{i-1}, D) \\ \leq \; & Q(\theta^{i}, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}|\theta^{i}), D) \\ = \; & l(\theta^{i}, D) \end{aligned}$$

Therefore, the algorithm obtains a better log likelihood after each iteration. Since the log likelihood is bounded, the algorithm will give an optimal configuration for alignments and paths of all the paired-end reads.

3.4 Implementation Details

Applying EM algorithm to millions of PERs to evaluate all their candidate alignments is computationally intensive. We have developed the following two strategies to speed up the computation.

3.4.1 Maximal exonic blocks

One of the most time consuming steps is the search for all possible splicing paths at each PER end read alignment. In the naive method, one may compute these paths one by one for each PER based on the splice graph G. However, such an implementation is not feasible for large RNA-seq data. To improve on it, our method first identifies the clusters of PER alignments sharing the same set of splicing paths. Therefore,

instead of computing the paths for one PER alignment at a time, the search can be conducted for the entire cluster.

The clusters were identified by partitioning the genome into maximum blocks in which no junction starts and/or ends. We define them as *m*aximum exonic blocks. To identify these blocks given a set of splice junctions, we start with the whole genome as one block. Each junction will be examined next. For each junction, if it falls into a block, the block will be split into two smaller blocks. For paired read alignment x, suppose its start read x_α is mapped to position $[a_1, b_1]$ and its end read x_β is mapped to position $[a_2, b_2]$. Then x_α can be mapped to a start block $B_\alpha = [B_\alpha^l, B_\alpha^r]$ and x_β can be mapped to an end block $B_\beta = [B_\beta^l, B_\beta^r]$, such that $B_\alpha^l \leq b_1 \leq B_\alpha^r$ and $B_\beta^l \leq a_2 \leq B_\beta^r$. After all the junctions are examined, the resulted blocks are all maximum blocks containing no junctions.

We then map all the paired end read alignments onto these blocks. If two paired read alignments π_{x_1} and π_{x_2} belong to the same start block and the same end block, they cover the same set of junctions and hence have the same set of possible paths. In this case, we group them into one *c*luster of PERs. For every cluster, we only need to compute the possible paths once. Then the particular set of possible distances for every alignment of this cluster can be calculated by adding the particular distance on the start block and the end block to the shared distance from the start block to the end block.

3.4.2 Independent Set of PERs

Performing iterative EM on all PERs is both memory and time consuming. Since most of the alternative splicing events occur locally within a gene and are independent among different genes, we adopt a divide and conquer approach by dividing the set of PERs into a number of *minimum independent sets*. Two sets of PERs are called independent if they do not share junctions. A set of PERs is a minimum independent set if it cannot be divided into two subsets of PERs that are independent. The probability of a PER is dependent on the junctions only if they overlap in their genomic span. This procedure helps to speed up the program significantly by confining EM procedures within each independent set, which is much smaller than the whole data.

3.5 Experimental Results

3.5.1 Improved splice junction detection on the breast cancer dataset

Datasets and parameters

We applied our methods on two 2x35bp paired-end RNA-seq datasets sampling two well-studied breast cancer cell lines, MCF-7 and SUM-102. The RNA-seq data were generated by the Illumina Genome Analyzer II.

Both datasets were first mapped by MapSplice by aligning all 35bp end reads individually. The error tolerance was set to 5%, allowing up to 2 mismatches in the alignment for each 35bp read. For spliced alignment, the minimum anchor size was 6 bp beyond the splice junction.

To understand how PERs might affect the sensitivity and specificity of junction

detection, no further filtering was performed on the alignments. Next, PER fragment alignment was computed using the methods proposed in this paper. The mate-pair distance distribution was fit to a Gaussian model with a mean of 112 bp and standard deviation of 40 bp.

The software was implemented in C++. The results presented here were run on an Intel(R) Xeon(R) E5540 (2.53GHz) CPU running Linux. The program is single-threaded and finished within 5 hours on each data set, using less than 10G memory. The software requires alignments of the individual end reads following the standard SAM format. In our case this was produced using MapSplice, but it can also be produced by TopHat or other RNA-Seq aligners producing read alignments in the SAM format. The output is the predicted alignment of the PER fragments also following the SAM format.

Table 3.1: Summary of the experimental datasets.

	same chromosome		cross-chromosome	
#PERs	input	mapped	input	mapped
MCF-7	12.7M	11.5M	541K	79K
SUM-102	13.6M	12.5M	527K	61K

Resolving Ambiguous Alignments

For each dataset, the number of input paired-end reads and the number of successfully mapped PERs are summarized in Table 3.1. About 91% of the PERs with both end reads mapped to the same chromosomes have fragment alignments with high probability. In contrast, less than 15% of the PERs have a highly probable fragment alignment if their end reads are mapped to different chromosomes. This might re-

flect the susceptibility of multiple alignment for short reads as a result of repeats or homologous genes across the genome.

Among the 11.5 million mapped PERs in the MCF-7 sample, about 7 million PERs have unique fragment alignments. Most of these PERs map onto exonic regions, and therefore contain only 49% of the junctions found among the single end reads. The rest of the mapped PERs either have ambiguous end alignments or ambiguous splicing paths. In these cases expectation maximization has assigned the most likely alignment. Without these alignments, it would be difficult to evaluate the quality of the majority of splice junctions. Restricting PER fragment alignments to unique alignments would also decrease junction coverage. The average support of the splice junctions covered by unambiguous alignments is 14.1 reads, whereas the average support from all PER alignments is 37.7 reads. Therefore, the expectation maximization method improves splice junction discovery as well as providing more accurate quantification of junction coverage, as shown in the next section.

Splice Junction Discovery

Sensitivity and specificity for splice junction detection. The alignment of the individual 35bp end reads yields a set J of putative splice junctions. During PER fragment alignment, the probability of each junction in J is evaluated according to the PER fragment alignments that incorporate it, and some putative junctions will be eliminated if there do not exist reliable PER alignments supporting them. We denote the set of junctions remaining following PER fragment alignment as J_{PER}. In Figure 3.5 (a) and (b) we compare the sensitivity and specificity of junctions detected

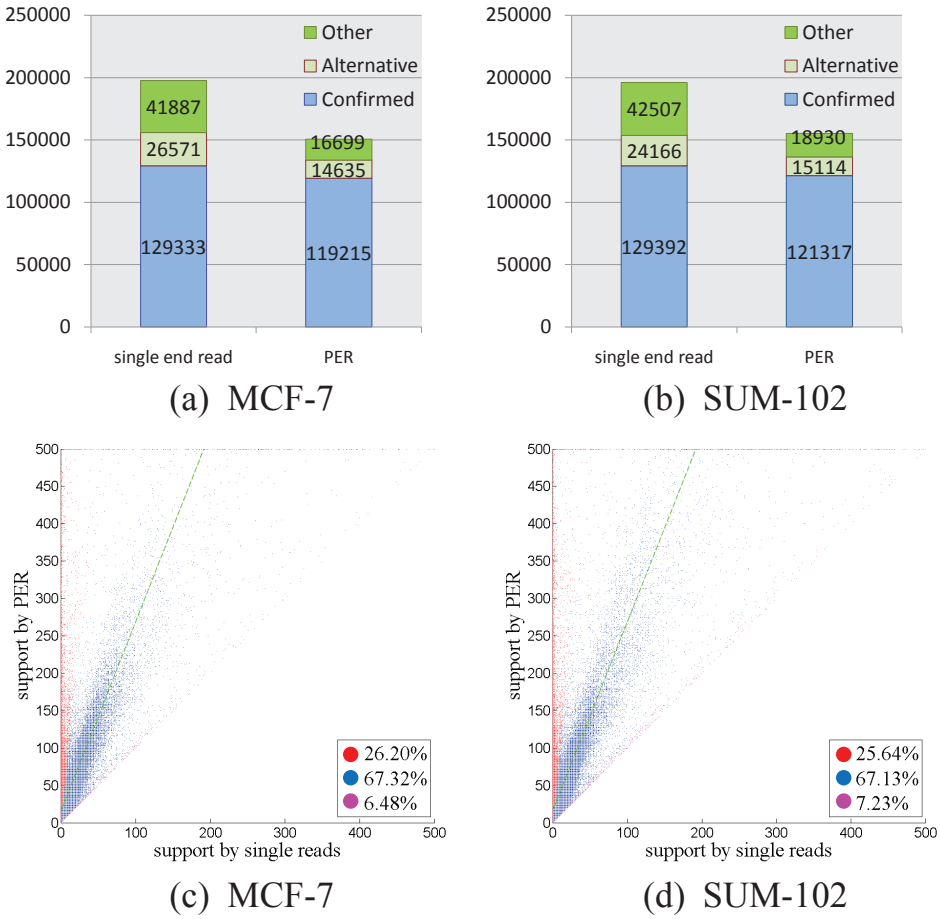


Figure 3.5: (a) and (b) *Comparison of sensitivity and specificity of splice junction discovery.* In each chart, the left bar represents junctions found by (spliced) alignment of PER end reads, and the right bar represents junctions found by alignment of the whole PER fragment. Each bar counts junctions in three categories: the bottom block is the number of junctions confirmed by GenBank; the middle block is the number of junctions whose 5' and 3' ends connect known exon boundaries or are close to such boundaries; the top block corresponds to the number of junctions that cannot be confirmed either way. (c)and (d) *Comparison of junction coverage.* For each confirmed junction, the x coordinate is the junction coverage among end read alignments, and the y coordinate is the junction coverage among PER fragment alignments. Points close to y axis, colored red, are junctions primarily supported by PER fragment alignments, while points close to the diagonal, colored magenta, are junctions primarily supported by end read alignments.

in end reads (J) with those remaining following PER fragment alignment (J_{PER}). In both datasets, about 79% of the junctions in J_{PER} were confirmed by transcripts in the Genbank database, while only 66% of junctions in J could be confirmed in this fashion. On the other hand, 93% of the total confirmed junctions in J were also present in J_{PER}. The small loss of sensitivity might be due to junctions present in one end of a PER whose other end failed to be aligned.

Among the unconfirmed junctions in J_{PER}, in both datasets nearly 50% were found to be either splice junctions connecting known exon boundaries or coordinates close to known exon boundaries. The majority of the unconfirmed junctions were highly supported and had coverage profiles resembling true junctions. In summary, splice junction discovery through PER fragment alignment mostly preserves the sensitivity of the discovery via individual end reads while significantly improving specificity.

Increased junction coverage with PER. We next look at how PER fragment alignment may change the coverage of junctions. The coverage of each junction $j \in J$ is the number of alignments of end reads that include j. Each $j \in J_{\mathrm{PER}}$ is covered by the number of PER fragments in which the junction is part of the most probable alignment. Since each PER fragment length is significantly longer than a single end read, we expect the coverage of junctions in J_{PER} to be significantly higher than the same junctions in J.

On both datasets, the average coverage of confirmed junctions is 37.7 using PER alignment, compared to only 11.1 using end read alignment. The scatter plots shown in Figure 3.5 (c) and (d) illustrate the PER support vs. single end support for all confirmed junctions. Around 25% of junctions are primarily supported by PER

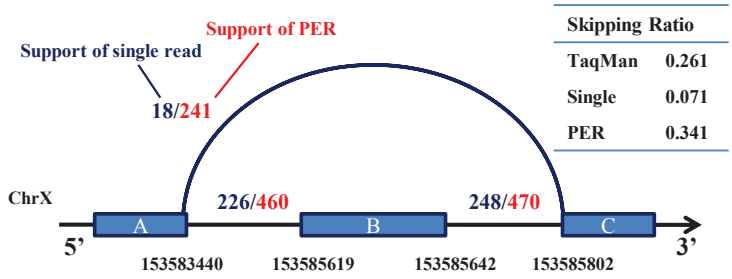
fragments, while only around 7% of junctions gain substantial support from single end reads. Furthermore, the majority of the junctions (more than 67%), corresponding to points colored blue, have PER support three fold higher than single end reads.

To evaluate the accuracy of the junction coverage in the presence of alternative splicing we selected 8 known skipped-exon alternative splicing events. We used quantitative RT-PCR to measure, in both of our data sets, the exon skipping ratio of the event, i.e. the fraction of transcript isoforms that include the preceding exon and the successor exon, but not the skipped exon. We compared these experimental values with exon skipping ratios calculated using the ratio of splice junction counts determined using individual end read alignments and using PER fragments alignments (Figure 3.6 (a)). With a Pearson correlation of 0.83 across all 16 measurements, the PER fragment alignments achieved high agreement with experimental values, as shown in Figure 3.6 (b). The accuracy is higher than the exon skipping ratio derived using counts from single end reads, which has a correlation of 0.78.

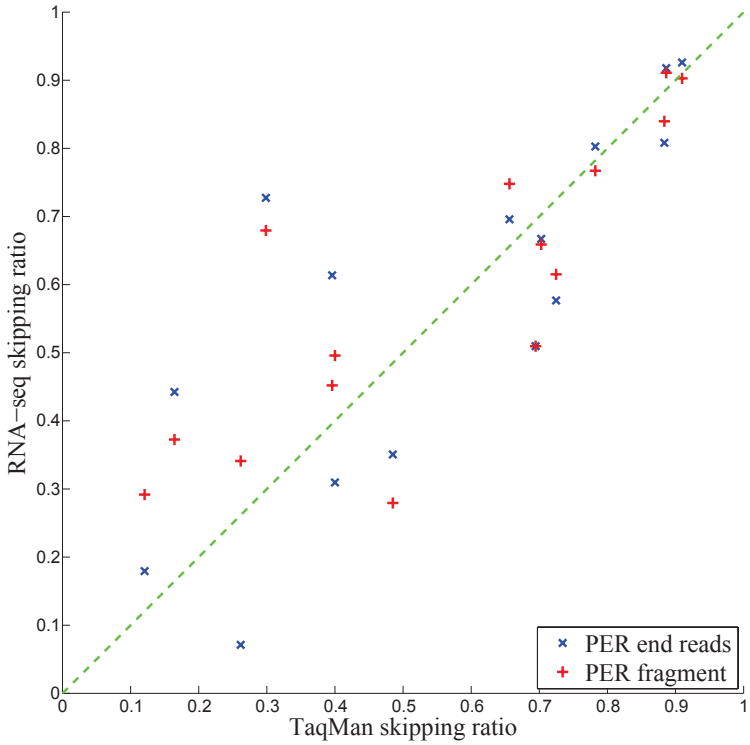
In summary, PER fragment alignment yields higher coverage of junctions than obtained from alignment of the end-reads only. The agreement with experimental measurements suggests that PER fragment alignment yields accurate coverage and assigns the correct splicing alternative to the individual PER fragment alignments that have splice graphs with alternative edges.

3.5.2 Comparison with SpliceMap

SpliceMap [Au et al., 2010] finds splice junctions in RNA-seq paired-end reads (PER). The algorithm performs independent spliced alignment of the end reads to a reference



(a)



comparison	correlation
Taqman - PER end reads	0.778955
Taqman - PER fragment	**0.828681**

(b)

Figure 3.6: (a) An example of an exon skipping event in gene FLNA with junction counts determined from the SUM102 RNA-seq data via end read alignments and PER fragment alignments, respectively. The skipping ratio is computed as $\text{count}(AC)/(\text{count}(AC) + \frac{1}{2}(\text{count}(AB) + \text{count}(BC)))$. (b) Correlation of 8 exon skipping ratios derived from qRT-PCR in each dataset and those computed using PER end read alignments and PER fragment alignments, respectively.

Table 3.2: Summary of the comparison results between SpliceMap and MapPER. The first column is the number of PER with endpoints aligned to the genome within the constraints of the respective methods. The second column records the fraction of end alignments that are valid alignments, i.e., consistent with an RNA fragment of size within the expected bounds from the RNA-seq protocol. The third column reports the number of junctions found within the end alignments, and the last two columns report the specificity of these junctions as measured by the ASTD database of known junctions (Koscielny et al., 2009).

	SpliceMap	MapPER
PERs Mapped	15.384M	16.964M
% with Fragment Alignment	%67.9	%100
Total Junctions	164,854	155,594
Confirmed Junctions	138,859	135,530
Specificity	%84.2	%87.1

genome and retains aligned pairs within 400,000bp in genomic coordinates. Splice junctions found in the ends of the retained pairs are the output of SpliceMap.

The MapPER algorithm described in this chapter has a different goal. Its output for a given paired-end read is the predicted alignment of the entire sampled RNA fragment, not just the sequenced ends or the junctions therein. It leverages the alignments of other reads in the same dataset to predict the alignment of the unsequenced portion of the fragment, guided by the expected fragment length that is part of the RNA-seq PER protocol.

To compare the two methods we compare the set of splice junctions found by each within the alignments of the PER ends and report how many of these junctions can be found in the ASTD database of junctions [Koscielny et al., 2009]. We also report how many of the endpoint alignments of each algorithm are consistent with the expected size of the fragments.

Experimental setup:

We applied both methods on the 2x50bp PER dataset used in the SpliceMap article [Au et al., 2010]. The data is publicly available in database GEO with accession number GSE19166. The protocol used to prepare the sample resulted in RNA fragments with a mean size of 114 bp between the sequenced ends and a standard deviation of 34 bp. The dataset contains about 23 million paired-end reads.

We applied *SpliceMap v3.2.1* to the dataset, tolerating up to two bp mismatch in each end alignment.

Our method used *MapSplice v1.12* with the same error tolerance to align reads to the reference genome and used *MapPER v0.1* to infer PER fragment alignments.

Since the SpliceMap method was designed to find only canonical junctions spanning less than 400K bp on the genome, we restricted MapSplice to the same conditions (MapSplice normally identifies non-canonical and fusion junctions as well).

PER fragment alignment: The MapPER fragment alignment algorithm is capable of resolving PERs with distances larger than normal range. We say a PER has a *valid alignment* if the inferred fragment alignment has a length in the expected range (within 3 s.d. of the mean). The results of SpliceMap method and MapPER method are compared in Table 3.2. MapPER aligned more PER than SpliceMap and all the alignments were required to be valid alignments. SpliceMap aligned end reads within 400K bp in genomic coordinates, among which 67.9% were classified as valid based on the expected mate-pair distances. The distances of the remaining 32.1% vary from 200bp to 400kb. Those alignments may harbor false positive alignments and false positive splice junctions. As a result, MapPER has higher specificity in terms of junction discovery than SpliceMap. It has similar sensitivity compared to

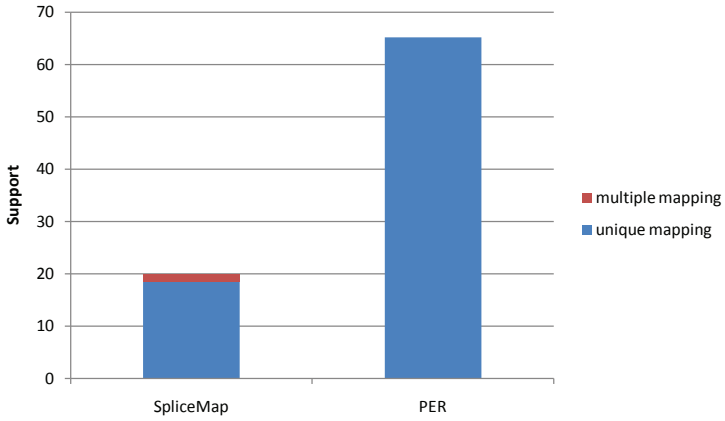


Figure 3.7: A comparison of the support of junctions discovered by SpliceMap and our method.

SpliceMap despite constraints for its accepted alignments that are more stringent.

Splice junction coverage: We compared coverage, the number of reads supporting a junction, derived from SpliceMap and our MapSplicePER method. As shown in Figure 3.7, the average junction support from SpliceMap is 19.9, including support of 18.5 from uniquely mapped reads and 1.4 from reads with multiple mappings. With PER fragment alignment, the support increases to 65.2, tripling the average junction support derived from SpliceMap.

In conclusion, the proposed MapPER method is able to perform fragment alignments for Paired-end RNA-seq reads. It increases specificity of the splice junction alignment without loss of sensitivity when the depth of sampling is high. Through the inference of the unsequenced fragments of PERs, it increases coverage of the transcriptome in general, and junctions in particular.

3.5.3 Consolidation of fusion junction discovery

Finally we apply the methods of this paper to the problem of gene fusion detection. Generally 35bp reads are too short to identify long range fusion junctions with any confidence, since genome-wide spliced alignment of a 35bp sequence will yield multiple occurrences due to chance as well as repeats and homologous genes.

We obtained a candidate set of fusion junctions from two sets of 75bp single read RNA-seq datasets from the same cell lines. The 75bp data set was aligned genome-wide using MapSplice without filtering (to maximize sensitivity) and candidate fusion junctions were selected by a spliced 75 bp alignment whose prefix and suffix (of length at least 25bp) were mapped to different genes.

Even by limiting the 75bp alignment to be unique, we obtained 13513 candidate fusion junctions in the MCF-7 dataset and 11665 putative junctions on SUM-102 dataset. Taking these fusion junctions as putative edges in the splice graph, our PER alignment using 2x35bp greatly reduced the possible fusion candidates. About 2904 junctions in MCF-7 and 2990 junctions in SUM-102 remained supported. This set of fusion junctions was further filtered by eliminating pairs of genes with high sequence similarity to avoid false positive predictions due to homologous genes. Figure 3.8 shows a final set of 18 fusion events where the genes connected by the junctions have less than 35% identity similarity evaluated by the Align program from Emboss. This includes 10 fusion events in MCF-7 and 8 fusion events in SUM102. Eight out of 10 MCF-7 fusion events were previously reported by Maher et al., 2009, where they were confirmed by experimental qRT-PCR validation. The detailed information of these

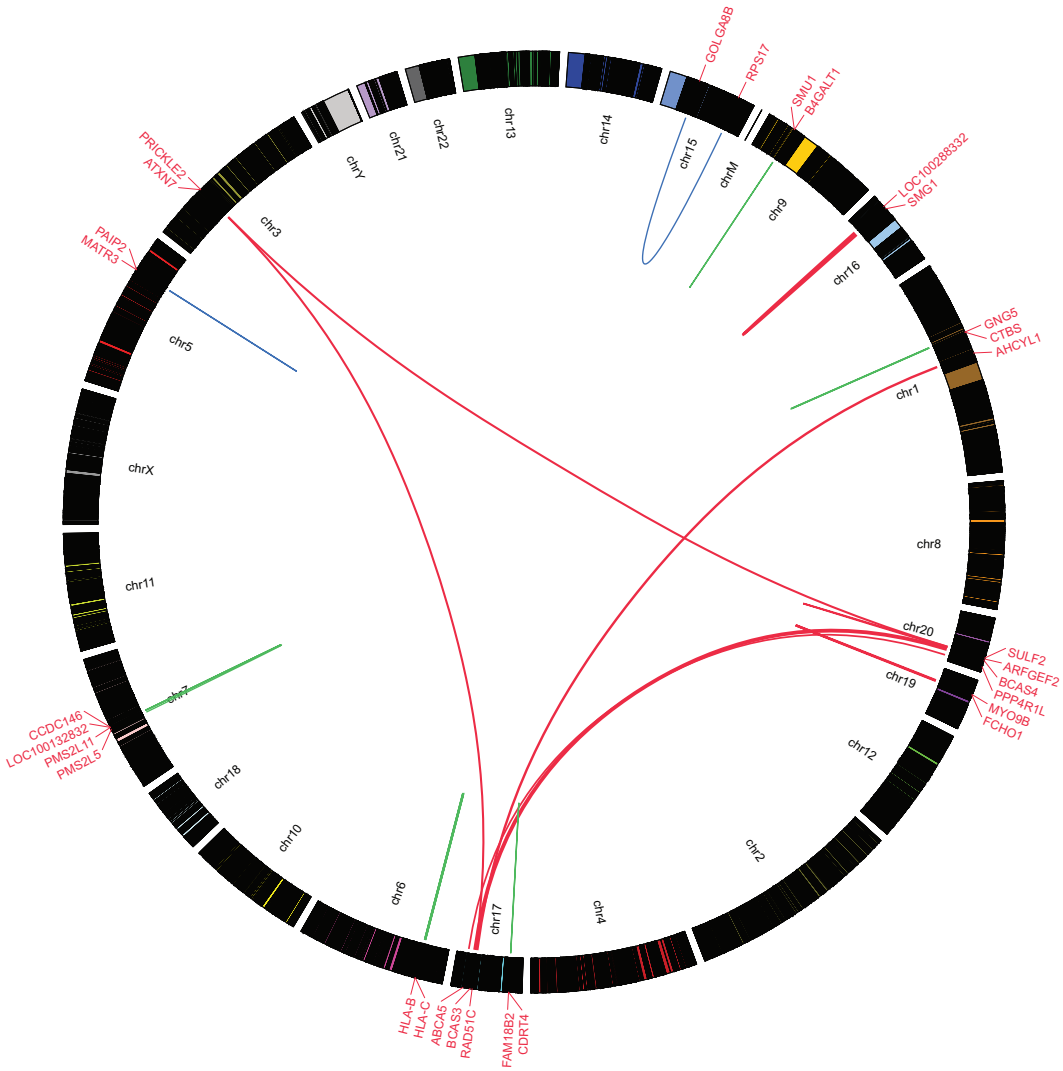


Figure 3.8: A set of gene fusion events confirmed by PER data, plotted with Circos [Krzywinski et al., 2009]. Red links refer to gene fusions events specific to MCF-7 alidated by Maher et al. [2009]. Blue links refer to two additional gene fusion events detected in MCF-7. Green links refer to the predicted gene fusion events in SUM-102.

gene fusion events are listed in Table 3.3.

3.6 Discussion

RNA sequencing using the paired-end protocol is a cost-efficient way to sample transcript fragments longer than the sequencing capability by sequencing only the ends.

Table 3.3: A list of rediscovered gene fusions specific to MCF-7 reported by Maher et al. [2009]. Most of the fusion junctions have much higher PER support than single 75bp read support.

Donor		Acceptor		Similarity	#PERs
BCAS4	chr20	BCAS3	chr17	24.3%	731
ARFGEF2	chr20	SULF2	chr20	27.6%	3
SULF2	chr20	PRICKLE2	chr3	29.7%	4
AHCYL1	chr1	RAD51C	chr17	22.4%	9
ATXN7	chr3	BCAS3	chr17	30.3%	4
LOC100288332	chr16	SMG1	chr16	3.3%	29
PPP4R1L	chr20	ABCA5	chr17	11.4%	3
MYO9B	chr19	FCHO1	chr19	28.7%	15

We propose a probabilistic framework to predict the alignment of each transcript fragment to a reference genome. The alignment chosen is determined by maximizing the likelihood of all PER alignments through an expectation maximization method.

PER transcript fragment alignment offers a number of advantages over the alignment of just the end reads. First, the fragment alignments significantly increase coverage of the transcriptome, providing a more robust measure of transcriptome expression profiles. Second, the splice junctions in the transcript fragments have higher specificity than the junctions in the individual end reads because the PER fragment alignments maximize information from the entire set of end read alignments. Third the splice graph accurately captures alternative paths between two end reads and the expected mate-pair distance of end reads can effectively disambiguate them, as shown by the high correlation with experimental measurement of alternative splicing events.

Another PER aligning method SpliceMap [Au et al., 2010] examines the PER support of a junction within some neighborhood and uses that to filter splice junctions.

However, lacking a splice graph model of the connection between the end reads, the method may miss true support and include spurious support especially in genes that are alternative spliced or are not highly expressed. In comparison, our likelihood-based method finds the accurate and complete set of PER supports without relying on an arbitrary threshold.

A major impetus for our work is the detection of novel gene fusion events that result from genomic rearrangement in cancer cells. However, identifying long range fusion junctions is particularly challenging due to the increased frequency of repeats and homologous genes at the genome wide scale. Our PER alignment approach is capable of detecting trans-chromosome and trans-strand gene fusion events, and the long length of the aligned transcript fragments make more likely the detection of such an event with highly significant long anchors on each side of the fusion. We have demonstrated the application of our method using 2x35bp PER reads together with single 75bp reads from MCF-7 and SUM-102 breast cancer cell lines. Our result detected 10 events, 8 of which are gene fusion events identified by Maher et al., 2009, demonstrating high specificity of the proposed method. If longer paired-end reads are used, such as 2x75bp, no additional single reads would be necessary for the initial fusion detection.

Copyright © Yin Hu, 2013.

Chapter 4 Genome-wide Detection of Alternative Splicing Events

4.1 Introduction

The mRNA transcriptome consists of all mRNA molecules transcribed from the genome within a functioning cell. Different genes give rise to different transcripts and may express differently. In addition, through the mechanism of alternative splicing, different subsets of exons in a gene may be concatenated (in transcription order) to form different transcript isoforms Sultan et al. [2008], Wang et al. [2008], Pan et al. [2008], Kwan et al. [2008]. The diversity and abundance of isoforms transcribed from a gene are known to vary in response to cellular differentiation and maturation as well as environmental factors and disease. The totality of transcripts present and their individual abundance characterizes the mRNA transcriptome and is a most basic phenotype. Thus the difference between transcriptomes sampled from healthy and diseased cells may provide insight into the functional consequences of disease, as well as identifying biomarkers to classify different disease types Wang and Cooper [2007]. Similarly, the difference between transcriptomes sampled at different stages in cell development may provide insight into the functional effects of cell differentiation and cell life cycles Wang et al. [2008], Trapnell et al. [2010].

Classically the differential analysis of transcriptomes has been studied using techniques such as microarray technologies Clark et al. [2002] that identify differences in the total expression of known gene transcripts and exon arrays Okoniewski and

Miller [2008], Xi et al. [2008] that detect differences in the expression of known gene exons. More recently, high-throughput sequencing methods such as RNA-seq Wang et al. [2009] have been able to accurately record short sequences of nucleotides sampled from millions of mRNA molecules in the transcriptome, and thereby are capable of observing samples from known as well as unknown transcripts, providing a more complete picture of the transcriptome. In addition, the large number of molecules sampled provide the potential to accurately estimate relative abundance of transcript isoforms.

Three basic strategies have emerged to identify *differential transcription*, the difference in the relative abundance of the individual transcripts across samples. The first strategy, *e.g.*, Cufflinks Trapnell et al. [2010], performs transcript inference and abundance estimation followed by differential test of relative abundance. Such an approach is ideal but its performance relies on accurate transcript quantification, which is itself a challenging problem. The RNA-seq reads generated by most sequencing platforms are less than 100nt single or paired end. In genes with a significant number of very similar alternative transcripts, they are too short to be assigned to individual transcripts unambiguously, making the transcript quantification problem underdetermined. Figure 4.1 demonstrates a gene with four isoforms as a result of two alternative splicing events. Transcripts could start and end at any exon, or even within exons. Assuming no transcript annotation is known, there can be more than one set of valid transcripts as shown in Figure 4.1b. Even with four known transcripts as given, there could be multiple solutions of valid quantification (Figure 4.1c). In this case, the problem of transcript quantification is *unidentifiable* Huang et al. [2012] and may result

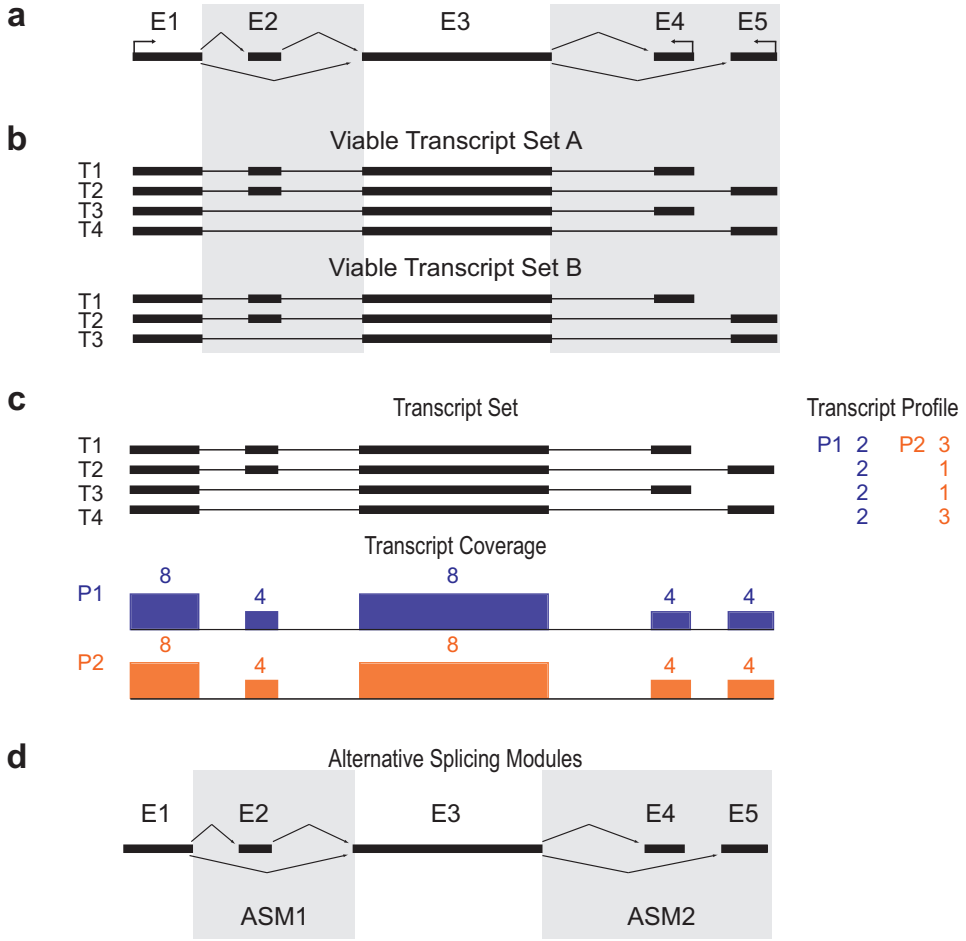


Figure 4.1: Challenges in using short reads to identify transcripts and their abundance. **(a)** An example of alternatively spliced gene with 5 exons. Each black rectangle denotes an exon and the arrows denote the splice junctions connecting the exons. Two alternative splicing events are present in this model: exon E_2 can be alternatively included or skipped by transcripts passing through E_1 and E_3, and transcripts passing through exon E_3 may alternatively end in E_4 or E_5. **(b)** Two viable transcript sets of the gene in (a), both explaining the splice variants suggested by the alternative splice junctions. **(c)** Two possible profiles of transcript abundance that have exactly the same expected coverage on exons and splice junctions, provided a transcript set. **(d)** The representation of alternative splicing with reduced complexity used by DiffSplice.

in inaccurate abundance estimation. Consequentially, the uncertainty in transcript quantification may lead to false discoveries of genes with differential transcription.

The second strategy indirectly detects differential transcription by aggregating

changes of multiple features on the transcriptome Stegle et al. [2010], Singh et al. [2011]. For example, a non-parametric statistical test called MMD (Maximum Mean Discrepancy) was designed in Stegle et al. [2010] for the comparison of read coverage on all exons. FDM (Flow Difference Metric) was designed to capture the average flow difference of all divergence nodes between two splice graphs Singh et al. [2011]. These approaches do not rely on any transcript information. However, they provide no simple localization of differences: MMD and FDM can only detect a diffuse "signal" of differential transcription without identifying the specific isoforms or regions that give rise to the difference.

The last strategy examines differential expression on annotated simple alternative transcription events in existing splicing databases. Examples include ALEXA-seq Griffith et al. [2010], MISO Katz et al. [2010], SpliceTrap Wu et al. [2011a], and MATS Shen et al. [2012]. These methods have been shown to be quite accurate in identifying differences in utilization of a skipped exon by isoforms in two samples. But they do not extend easily to more complex alternative splicing patterns with more than 2 alternative splice forms. These methods cannot be easily generalized to accommodate novel alternative splicing events that can be discovered by RNA-seq data, consequentially misinterpreting the data and the splicing events.

In this chapter, we present an *ab initio* method named DiffSplice for the detection and visualization of differential alternative transcription. DiffSplice circumvents the need for full-length transcript inference and quantification and localizes its search at Alternative Splicing Modules (ASMs) (Figure 4.1d). These modules represent the genomic regions where alternative transcripts diverge, localizing the nature of the

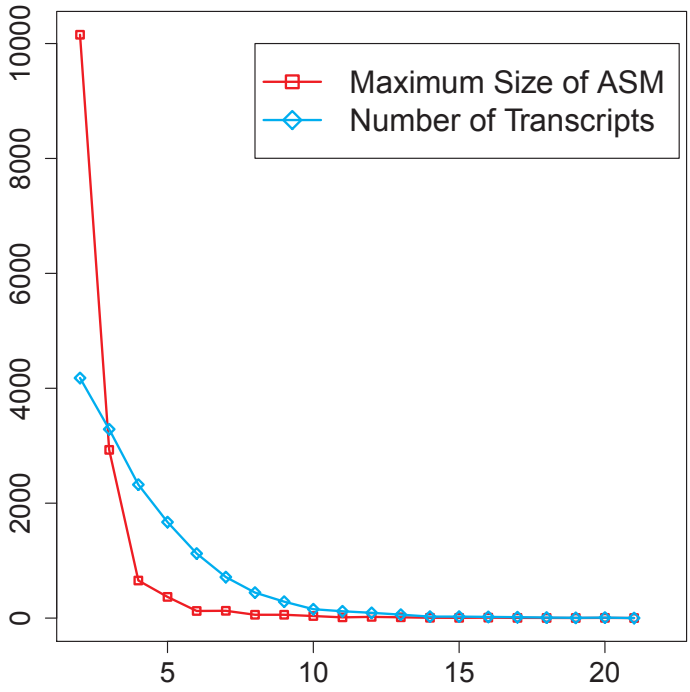


Figure 4.2: The distribution of the number of alternative transcripts per gene with UCSC human hg19 RefSeq annotation and the number of alternative paths per ASM after decomposition. The plot shows ASMs have significantly fewer alternative paths. The reduced complexity allows more accurate quantification.

difference and decreasing the complexity of the differential analysis by comparing corresponding ASMs between samples (Figure 4.2). The ASMs are detected automatically from a transcriptome-wide expression-weighted splice graph (ESG), which is built directly from read alignments and captures all the sample-relevant splicing events including novel ones. Expression estimation of associated isoforms and tests for differential transcription start from the simplest ASMs, which yields estimation that is more robust to sequencing bias, and work outward. A non-parametric statistical test is introduced to assign the significance level of the differential transcription in the ASMs with a controlled false discovery rate. By design, differential analysis on

ASM can be performed using short reads.

Our results on synthetic datasets demonstrate the precision of DiffSplice in the discovery and the expression estimation of ASMs and hence the sensitivity in the quantitation of transcriptional differences between samples. Simulation experiments on human transcriptome support the robustness of our method at different sampling depths and under various sampling biases. We applied DiffSplice on a time course lung differentiation dataset, where 498 genes were tested to have significant change of transcription as well as 2077 with significant change of overall gene expression, supporting the hypothesis that differential transcription is the key in the mucociliary cell differentiation and function. We also discovered 910 novel alternative splicing events that were not present in existing RefSeq and UCSC transcript annotations. The consideration of replicates in test statistics allowed DiffSplice to account for sample variations, reducing the risk of unreliable discoveries. Beyond the scope of differential transcription in alternatively spliced exons, the application of the proposed method on a breast cancer dataset discovered cell line specific structural variations such as deletions, demonstrating the feasibility in identifying irregular transcription variants that may reveal crucial regulatory mechanism in a cancer transcriptome.

4.2 Related work

4.2.1 Differential transcription analyses

As discussed in Section 3.1, in general philosophy the existing methods for differential transcription analysis may be divided into 4 basic strategies: the gene-level

Table 4.1: Methodological comparison of differential transcription analyzing approaches using RNA-seq — transcription event reconstruction and quantification.

Category	**Method**	**Description**
Granularity of differential analyses		
Gene level	MMD, FDM	Compare overall read distribution on features in the gene, such as the exons and the alternative splice sites
Transcript level	Cufflinks, combinations of transcript quantification approaches and differential gene expression analyzing methods	Directly estimate and compare the abundance of transcripts in the transcriptome
Alternative splicing level	Alexa-seq, MISO, SpliceTrap, MATS, DiffSplice	Estimate and compare the abundance of splicing variants at alternative splicing events in the transcriptome
Transcription specification		
Reference transcriptome-based	MMD, RSEM, Alexa-seq, MISO, SpliceTrap, MATS	Rely on transcriptome annotation to generate the gene model, transcript set or alternative splicing events
Ab initio, annotation-free	FDM, Cufflinks, DiffSplice	Use only read alignments the reference genome to reconstruct splice graph, transcripts or alternative splicing events
Variant expression estimation/interpretation		
Observed read distribution	MMD, FDM	Use the observed read counts or the number of junction spanning reads directly with no estimation
Exon skipping ratio	Alexa-seq, MISO, SpliceTrap, MATS	Estimate the ratio of exon skipping variant against exon retaining variant, allowing only 2 variants typically
Alternative path abundance	Cufflinks, DiffSplice	Estimate the relative proportions of all alternative transcripts/splicing paths, allowing more than 2 variants

Table 4.2: Methodological comparison of differential transcription analyzing approaches using RNA-seq — differential analysis.

Method	**Description**
Test statistic of differential transcription	
Cufflinks	FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for transcript expression level, the Jensen-Shannon divergence between relative transcript proportions of a gene at different conditions
MISO	The Bayes factor comparing hypothesis of differential transcription against hypothesis of non-differential transcription
FDM	The averaged flow difference at all splice sites
DiffSplice	Read coverage for splicing variant expression level, the relative difference statistic measuring the ratio of between-group discrepancy against within-group variance
Assessment of statistical significance of differential transcription	
Cufflinks	The p-value according to the asymptotic distribution of Jensen-Shannon divergence, derived from the Delta's method
MISO	The magnitude of the Bayes factor
MATS	The posterior p-value of inclusion-ratio change, evaluated using MCMC (Markov Chain Monte Carlo)
FDM	The p-value from permutation test by permuting reads
DiffSplice	The false discovery rate over all genes from permutation test by permuting samples

diffuse analysis such as MMD [Stegle et al., 2010] and FDM [Singh et al., 2011], the transcript-level differential test like Cufflinks [Trapnell et al., 2010], the annotation-based exon inclusion/exclusion analysis such as ALEXA-seq [Griffith et al., 2010], MISO [Katz et al., 2010], SpliceTrap [Wu et al., 2011a] and MATS [Shen et al., 2012], the data-driven differential splicing event analysis like DiffSplice [Hu et al., 2012].

The other detailed differences among the methodologies are summarized below

and in Table 4.1 and Table 4.2.

Granularity of differential test Methods perform differential transcription analysis in different unit. The indirect methods, such as MMD and FDM, collect expression measurements in the basis of gene, without knowledge of transcripts. The transcript-based methods, such as Cufflinks, estimate abundance of transcripts from annotation or transcript reconstruction procedures and test for differential expression in the basis of transcript. The last category, including ALEXA-seq, MISO, SpliceTrap, MATS and DiffSplice, localizes differences at alternative splicing events. The signal of differential transcription may have different magnitude when looking at different levels of a gene and hence may cause disagreement on results of differential test.

Alternative variants specification Methods using or not using transcript annotation interpret the data with different set of splicing variants/transcripts and thereafter may have different conclusions on many genes. Methods depending on transcript annotation, including MMD, ALEXA-seq, MISO, SpliceTrap and MATS, are challenged by the uncertainty of the existence of the annotated transcripts and the incompleteness of the transcript set in the annotation; Methods utilizing only RNA-seq read alignments, such as Cufflinks, FDM and DiffSplice, are challenged by the high ambiguity between actually expressed loci/splicings and noise.

Variant expression estimation The expression estimation methods are deviated. MMD uses the observed read counts on every base; FDM uses the observed spliced read counts on every junction; MATS estimates the exon skipping ratio based on the spliced read counts on the skipping and retaining junctions; MISO and Splice-

Trap estimates the exon skipping ratio using a Bayesian technique; Cufflinks and DiffSplice estimate transcript abundance and alternative path abundance following a Poisson model. The accuracy of the estimated variant expression directly determines the correctness of differential test statistics.

Differential test statistic Differences in splicing/transcript variants' expression between sample groups are summarized into differential test statistics. For example, Cufflinks calculates the Jensen-Shannon divergence (JSD) that compares the relative proportion of transcripts in a gene. MISO presents the Bayes factor that calculates the marginal likelihood ratio of the differential transcription model against the non-differential transcription model. FDM averages the difference in the proportion of alternative splicing variants at every splice site in a gene. DiffSplice incorporates the JSD and the SAM statistic [Tusher et al., 2001] to develop a test statistic that directly measures the discrepancy between group-wise transcript proportion profiles, taking into account the within-group variance and the expression level.

Assessment of significance Last but not least, the statistical tests that evaluate the significance of the statistics vary largely. Cufflinks derives the asymptotic distribution of JSD using the Delta method and calculates the p-value of the observed difference on each gene according to the asymptotic distribution. MATS uses the Markov chain Monte Carlo method to assess the posterior p-value of the change on exon-inclusion ratio for every exon skipping event. FDM and DiffSplice both use permutation test, but they test different things. For two groups of samples, FDM compares all pairs of samples and selects genes with large difference of the number of between-group significances and the number of within-group significances. Every

gene is tested separately. In comparison, DiffSplice selects significant differences by comparing the calculated statistics to the expected statistics on all alternative splicing modules over the transcriptome, not only on a single gene, and controls false positives with the false discovery rate.

4.3 Construction of expression-weighted splice graph

4.3.1 Construction of Transcriptome-wide unified Expression-weighted Splice Graph (ESG)

Traditionally, the transcriptome is either represented by a list of transcripts Trapnell et al. [2010] or a splice graph Heber et al. [2002], Hu et al. [2010], Singh et al. [2011]. In comparison, a list of individual transcripts encodes the complete set of transcriptional information, whereas a splice graph summarizes the variation among multiple transcripts and clearly shows the exons that may be spliced out during transcription as well as the exons that are always retained. With RNA-seq reads, the prediction of individual exons and splice junctions has become a routine, allowing accurate reconstruction of the splice graph. The prediction of full-length mRNA transcripts remains challenging especially for genes with highly complex alternative splicing events. Therefore, our method starts with the construction of a splice graph.

The splice graph is built from the RNA-seq read alignments to the reference genome. Alternatively, it can be built *de novo* by assembly of RNA-seq reads Birol et al. [2009], Grabherr et al. [2011], Li et al. [2010c]. The alignment of RNA-seq reads to a reference genome has been studied extensively in the past two years Wang et al.

[2010a], Trapnell et al. [2009a], Au et al. [2010]. There exist two types of read alignments, exonic alignments and spliced alignments. An exonic alignment corresponds to a contiguous sequence of nucleotides on the genome, typically indicating expressed exonic regions. A spliced alignment spans two or more exons, consequentially defining the donor and acceptor sites of the splice junctions. For paired-end reads (PER), DiffSplice first applies MapPER Hu et al. [2010] to determine the whole transcript fragment alignments according to the distribution of the expected mate-pair distance (Figure 4.3a), which allows more accurate splice prediction and expression profiling.

In a splice graph $G = \langle V, E, w \rangle$, every node corresponds to an exonic unit, an expressed region on the genome whose boundaries are delimited by donor and acceptor splice sites defined by the location of splice junctions. It is difficult to detect the precise transcription start and end sites with RNA-seq reads from commonly used library prep protocols. They are therefore estimated as the locations where read coverage changes significantly from absence to presence and vice versa relative to background, respectively. With alternative splice sites, part of an exon can be skipped in one transcript but not in others. In this case, a continuous exonic region will be further divided into smaller units, allowing each of them to be alternatively included in transcripts. Since the exonic units are linearly ordered on the reference genome, nodes in V can be ordered based on their locations on the genome. We say $v_s < v_e$ if the location of v_s is upstream of v_e in the direction of transcription. Two exonic units will be connected by an edge if there exist read alignments that contiguously cover both of them. The direction of the edge is determined by the direction of the transcription identified by the dinucleotide sequences in the intron flanking the

donor and acceptor sites. For example, a GT-AG dinucleotide pair flanking the intron sequences in the reference genome suggests forward transcription, while the CT-AC pair suggests the reverse transcription. The expression levels on the exonic units and the splice junctions are then collected as the weights w of the vertices and the edges.

To make the description of the following algorithm easier, we further augment the general splice graph $G = \langle V, E, w \rangle$ by adding a virtual transcription start node ts and a virtual transcription end node te. Edges will be added to connect the start node ts to all the vertices where transcripts initiate and similarly to connect all the vertices where transcripts terminate to the end node te. Therefore, all transcripts in a gene will start from ts and end in te. We also assume for every vertex $v \in V$ there is a directed path from ts to v and a directed path from v to te, that is, every exonic segment can be reached by some transcript in the gene. We refer to the augmented splice graph as the Expression-weighted Splice Graph (ESG).

4.4 Identification of differentially transcribed loci

4.4.1 Detection of Alternative Splicing Module (ASM)

Next we identify alternative exonic events through the decomposition of the ESG into *alternative splicing modules (ASMs)*. An ASM is defined as a single-entry and single-exit subgraph of the splice graph. The entry node is the only exonic unit where transcripts can flow into the ASM; Similarly, the exit node is the only node where transcripts leave the ASM. Transcripts diverge into more than one isoforms by following different paths in the ASM before reconvening at the exit node.

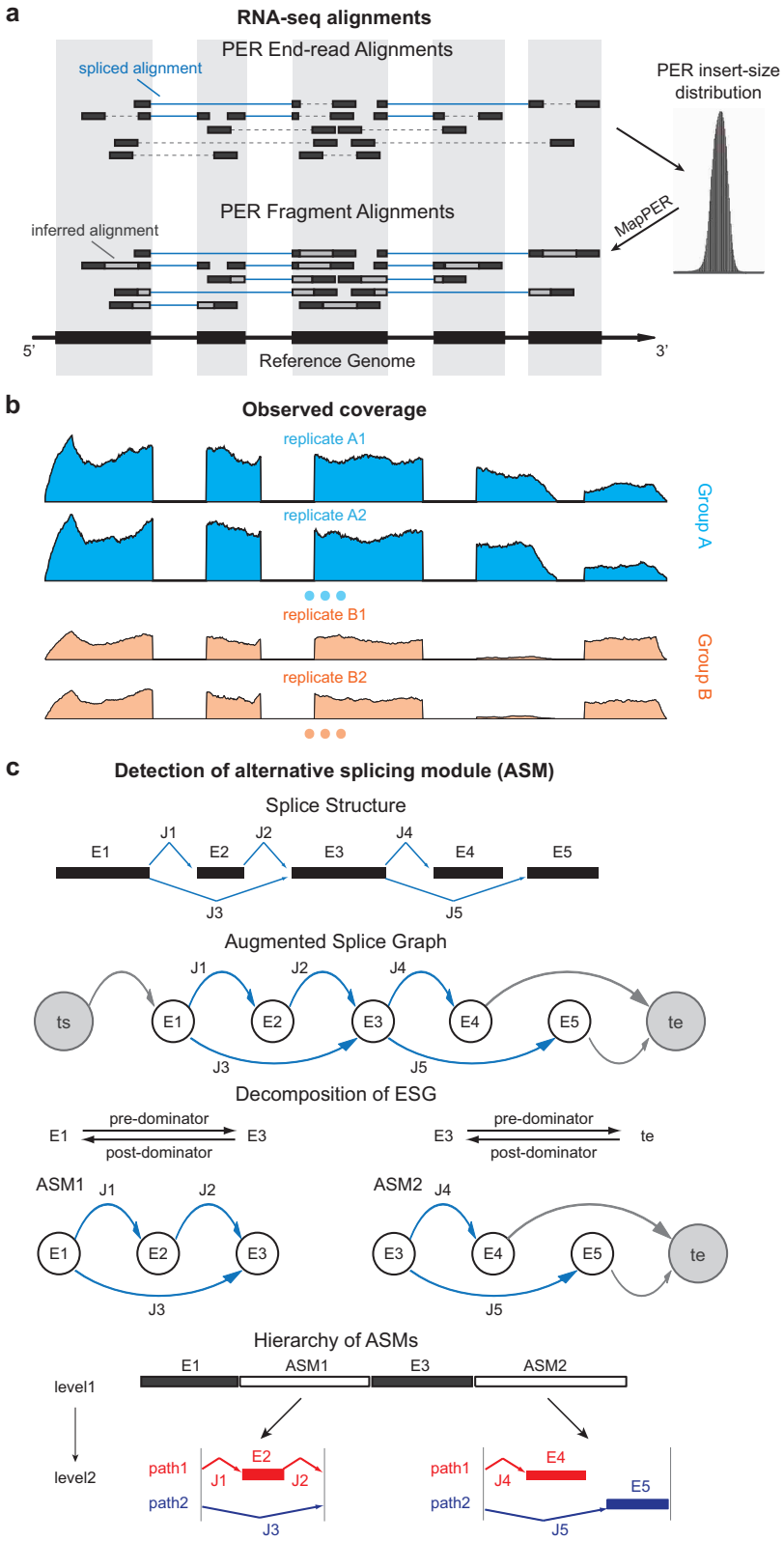


Figure 4.3: Overview of the DiffSplice framework — construction of the genome-wide unified expression-weighted splice graph and identification of the alternative splicing modules

Let $G = \langle V, E, ts, te, w \rangle$ be the ESG of a gene. A vertex $u \in V$ *pre-dominates* a vertex $v \in V$ if every path from the transcription start ts to v (include v) contains u. A vertex $w \in V$ *post-dominates* a vertex $v \in V$ if every path from v to the transcription end te (include v) contains w. Additionally, u/w is the *immediate* pre/post-dominator of v if every other vertex $x \in V$ that pre/post-dominates v also dominates u/w. We define the out-degree and the in-degree of a vertex $v \in V$ as the number of out-going edges and the number of in-coming edges of v, denoted as $d^+(v)$ and $d^-(v)$, respectively.

Definition A subgraph H of a graph G is said to be *induced* if, for any pair of vertices x and y of H, xy is an edge of H if and only if xy is an edge of G. In other words, H is an induced subgraph of G if it has exactly the edges that appear in G over the same vertex set. If the vertex set of H is the subset S of V(G), then H can be written as G[S] and is said to be induced by S.

Definition An ASM is an induced subgraph $H(ts_H, te_H) =< V_H, E_H, ts_H, te_H >$ of G with a distinguished node ts_H not in H as the *entry* and a distinguished node te_H not in H as the *exit* satisfying the following conditions

1. *Single-entry*: all edges from $(G - H)$ to H come from ts_H;
2. *Single-exit*: all edges from H to $(G - H)$ go to te_H;
3. *Alternative paths*: $d^+(ts_H) > 1$ and $d^-(te_H) > 1$;
4. *Minimal*: there does not exist a vertex $v \in V_H$ such that v post-dominates ts_H or pre-dominates te_H in $H(ts_H, te_H)$.

Having an ASM being single-entry and single-exit makes it an independent observation of the transcriptome. The number of transcript copies that go through an ASM can be entirely determined by the number of transcript copies passing through the entry node and exit node. There does not exist additional flow of transcripts. This property allows robust local abundance estimation within each ASM.

One ASM might be *nested* within another ASM if it is a subgraph of the bigger one. For two distinct ASMs $H_1(ts_1, te_1)$ and $H_2(ts_2, te_2)$, H_2 is nested in H_1 if and only if ts_1 pre-dominates ts_2 and te_1 post-dominates te_2. If there exists no H_3 such that H_2 is nested in H_3 and H_3 is nested in H_1, we say H_2 is a *child* of H_1 and H_1 is the *parent* of H_2. The parenting and nesting relation among the ASMs will form a hierarchy, showing how transcripts in the gene diverge and reconvene from transcription start sites to transcription end sites. Figure 4.4 shows the hierarchical decomposition on gene VEGFA. A total of 6 ASMs result from the decomposition. In the resulting hierarchy, if H_1 is an ancestor of H_2 (*i.e.*, H_2 is nested in H_1), the transcripts flowing into H_2 must be a subset of the transcripts in H_1. If H_1 and H_2 have the same parent (*i.e.*, H_1 and H_2 are siblings) and are on the same path, the transcripts passing through H_1 and H_2 are the same and the expected expression of H_1 and H_2 are the same.

Here we outline the algorithm that decomposes an ESG $G =< V, E, ts, te, w >$ into a set of ASMs. The pseudo-code can be found in Supplementary Section 1. Step 1-2 describe the procedure to determine ASMs within an ASM-type subgraph, and step 3 decomposes the subgraph which allows the iterative identification of all ASMs in the gene. To initialize, we start with the entire ESG G.

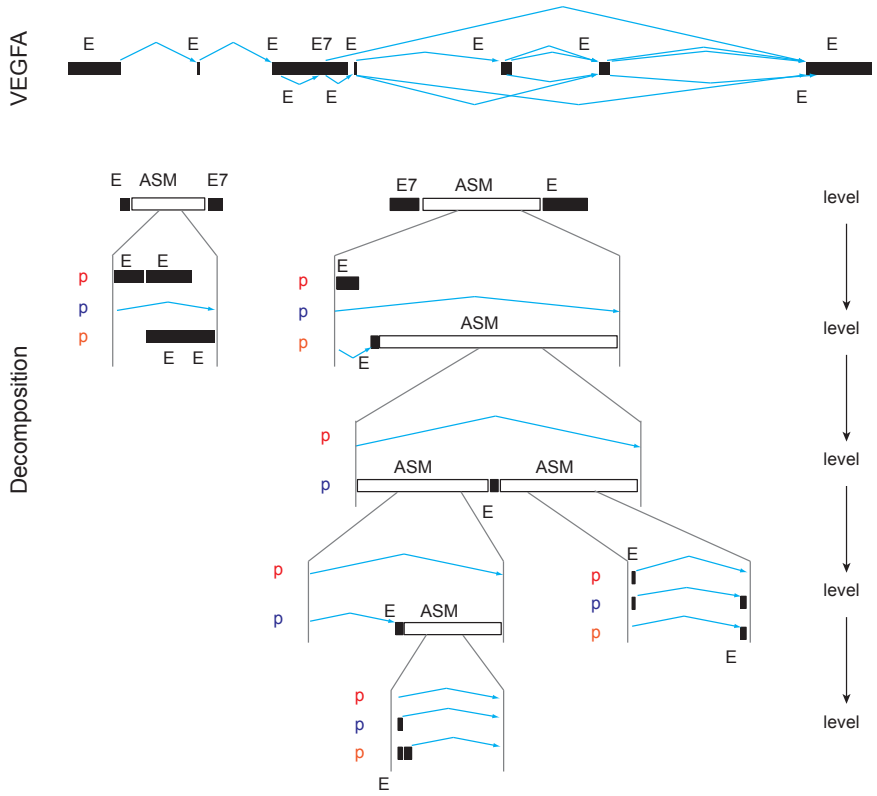


Figure 4.4: The splice graph and the ASM decomposition of gene VEGFA.

Step 1. Calculate the immediate pre/post-dominators. We first calculate the immediate pre-dominators and post-dominators of every vertex $v \in V$. The pre-dominators for vertex v (other than v) can be found by iteratively intersecting the sets of pre-dominators for all predecessors of v Ferrante et al. [1987], Pingali and Bilardi [1997]. Similarly, the set of post-dominators for v is the union of v and the intersection over the sets of post-dominators for all successors of v. According to the approach proposed in Buchsbaum et al. [1998], the bottom nodes of the depth-first search tree of G are grouped a collection of small, vertex-disjoint regions called *microtrees*. For vertex v, the aforementioned union-intersection operations are then performed locally within the microtree where the immediate dominator of v resides.

Step 2. Discover ASM. Candidate entries or exits for ASMs are the vertices with out-degree or in-degree larger than 1. Let u and v be two vertices in V such that $d^+(u) > 1$ and $d^-(v) > 1$. If u pre-dominates v and v post-dominates u and there does not exist a third vertex $w \in V$ such that u pre-dominates w and v post-dominates w, the subgraph bounded by u and v, denoted as $H(u, v)$, forms an ASM.

Step 3. Discover nested ASM. For any two edges (u, v) and (u', v'). We order $(u, v) > (u', v')$ if and only if there exists a directed path from u to u' and a directed path from v' to v. Hence the edges in H form a partial order. If there is no edge (u'', v'') in H such that $(u'', v'') > (u, v)$, edge (u, v) is called a maximal edge. We remove all the maximal edges in H and iteratively go to step 1 to resolve all nested ASMs until no new ASMs can be found in step 2.

Following is the pseudo-code for the algorithm to decompose an ESG.

```
input  : G =< V, E, ts, te, w >
output: Pre − Dom(v) for every v ∈ V
Pre − Dom(ts) ← {ts};
for v ∈ V do
    Pre − Dom(v) ← V ∪ {ts};
end
while changes in any Pre − Dom(v) do
    for v ∈ V do
        for u is a predecessor of v do
            Pre − Dom(v) ← {v} ∪ Pre − Dom(v) ∩ Pre − Dom(u);
        end
    end
end
```

Algorithm 1: Find Pre-dominators

The time complexity of the first step is linear in the number of vertices and edges Buchsbaum et al. [1998], or $O(|V|+|E|)$. In the second step, for every candidate

```
input  : G =< V, E, ts, te, w >
output: Post − Dom(v) for every v ∈ V
Post − Dom(ts) ← {te};
for v ∈ V do
    Post − Dom(v) ← V ∪ {te};
end
while changes in any Post − Dom(v) do
    for v ∈ V do
        for u is a successor of v do
            Post − Dom(v) ← {v} ∪ Post − Dom(v) ∩ Post − Dom(u);
        end
    end
end
```

Algorithm 2: Find Post-dominators

```
input  : G =< V, E, ts, te, w >
output: E_max
E_max ← any edge {e ∈ E};
for all e_1 = (u_1, v_1) ∈ E do
    for all e_2 = (u_2, v_2) ∈ E_max do
        if there is a path from u_1 to u_2 and a path from v_2 to v_1 then
            E_max ← E_max \ {e_2} ∪ {e_1};
        end
    end
end
```

Algorithm 3: Find maximal edges in an ESG G (*CalculateMaximalEdges(G)*)

entry the search of its paired ASM exit checks whether its immediate post-dominator is a candidate exit and also immediately pre-dominated by the entry, taking time of $O(|V|)$. In the last step, the maximal edges according to the partial order can be selected by iterating over all edges in E and keeping track of the maximal edges, resulting in an $O(c|E|)$ time scheme. Here c denotes the number of maximal edges in G. Because c is typically very small in a splice graph, the time complexity of the third step can be viewed as $O(|E|)$ in our application. Therefore, the time complexity of identifying ASMs from an ESG G is $O(|V| + |E|)$ and the time for discovering all

input : An ESG $G =< V, E, ts, te, w >$, parent Δ_P
output: The set of ASMs $\mathcal{A}$
Calculate pre-dominators in G;
Calculate post-dominators in G;
$Candidate_{entry} \leftarrow \{u : d^+(u) > 1\}$;
$Candidate_{exit} \leftarrow \{v : d^+(v) > 1\}$;
for *all* $u \in Candidate_{entry}$ **do**
  $v \leftarrow$ the immediate post-dominator of u;
  if $v \in Candidate_{exit}$ *and* u *is the immediate pre-dominator of* v **then**
    $parent(H(u, v)) \leftarrow \Delta_P$;
    $\mathcal{A} \leftarrow \mathcal{A} \cup H(u, v)$;
    $E_{max} \leftarrow CalculateMaximalEdges(H(u, v))$;
    $Decompose(H(u, v) \backslash E_{max}, H(u, v))$;
  end
end

Algorithm 4: Find the ASMs in an ESG G (*Decompose(G,*Δ_P*)*)

nested ASMs is dependent of the total number of ASMs.

4.5 Biological applications of ASM

In practice, ASM highlights the region(s) of a gene that vary among isoforms. Ultimately, the biologist needs to know how differences among isoforms result in a change in the biological activity of the protein they encode. Making the leap from RNA-seq data to protein sequences encoded is a challenge for all approaches using RNA-seq data to detect differential transcription. While it is possible to infer the alternative isoforms from RNA-seq data (e.g. Cufflinks), these inferred transcripts are often inaccurate and can lead to mistaken conclusions. In DiffSplice, we do reconstruct subsections of transcripts and estimate their abundance they are just not the full length transcripts but the fractions that distinguish the transcripts. Instead we focus on the parts of the gene that are variable as a result of treatment or condition,

rather than those that are unchanged. The output of DiffSplice highlights not only regions with significant differences in transcription, but also the reconstructed transcript pieces that constitute the region together with their estimated abundance in different conditions. Furthermore, the local events identified by DiffSplice are meaningful in the concept of transcripts. When a gene has only one ASM, its transcripts have one-to-one correspondence towards the paths in the ASM. That genes transcripts are differentially transcribed if and only if one or more paths of the ASM are differentially expressed. When a gene has more than one ASM, one path of an ASM might relate to more than one transcript. However, any differential expression detected in ASM level will allow us to focus on only a small subset of candidate isoforms for further validation. A routine qPCR is often sufficient to confirm that the most abundant transcript does encompass the most abundant edges of two or more ASMs. Moreover, investigation of the functional and structural consequences of the alternate exons may reveal what is occurring in genes with multiple ASMs. Do the exons in an ASM add or remove a known motif? Does that motif require changes in another part of the protein that is reflected in the other ASM? Are the conserved regions the same among samples? Do nested ASMs correspond to hypervariable regions of the protein? In short, careful annotation of an ASM will typically reveal as much as an inferred isoform with little to no risk of an inaccurate reconstruction. For the problem of differential transcription detection, this analysis in the unit of ASM has much improved sensitivity and specificity than existing approaches based on full length transcript reconstruction and quantification. Ours is a more fine grained approach than the whole transcript view and can potentially identify rare cases where

mean expression of the entire gene is the same, but there are dramatic differences in the isoforms produced.

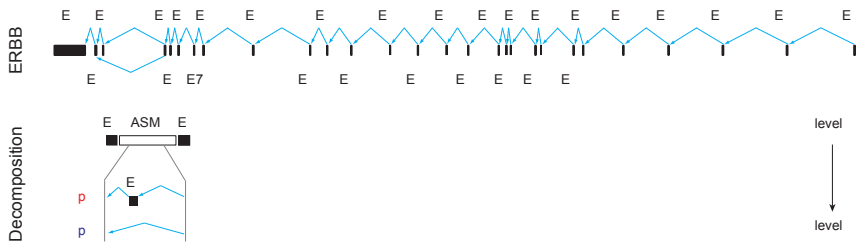


Figure 4.5: The splice graph and the ASM decomposition of gene ERBB4.

Here we give three examples to demonstrate that the investigation of ASMs may reveal functional sequences. The first two examples (ERBB4 and VEGFA) show significant sequences residing in single ASMs, while the third example (CD44) show an isoform transition associated with multiple ASMs.

In Figure 4.5 we plot the ASM in gene ERBB4. ASM_1 indicates an exon skipping event that alternatively includes or excludes exon E_3. The skipping path (p_2), which corresponds to the CYT-2 isoform in ERBB4, deletes a WW binding motif, leading to increased cell proliferation. Muraoka-Cook et al. [2009]

We take gene VEGFA as another example which has 6 ASMs with complex nesting structure. Bainbridge *et al.* have identified a 7-amino acid peptide, RKRKKSR, encoded by exon E_{10}. Bainbridge et al. [2003] This peptide could inhibit VEGF receptor binding and angiogenesis in vitro. In Figure 4.4 we show the ASMs in gene VEGFA. $ASM3.1$ captures the alternative inclusion/exclusion of E_{10}. Thus, this ASM shows that some isoforms of VEGFA lack this important peptide sequence.

Lastly, we look at two isoforms in gene CD44, CD44s and CD44v. Isoform CD44s includes exons $E_1 - E_5$, $E_{14} - E_{17}$ and E_{18}, and CD44v includes exons $E_1 - E_5$,

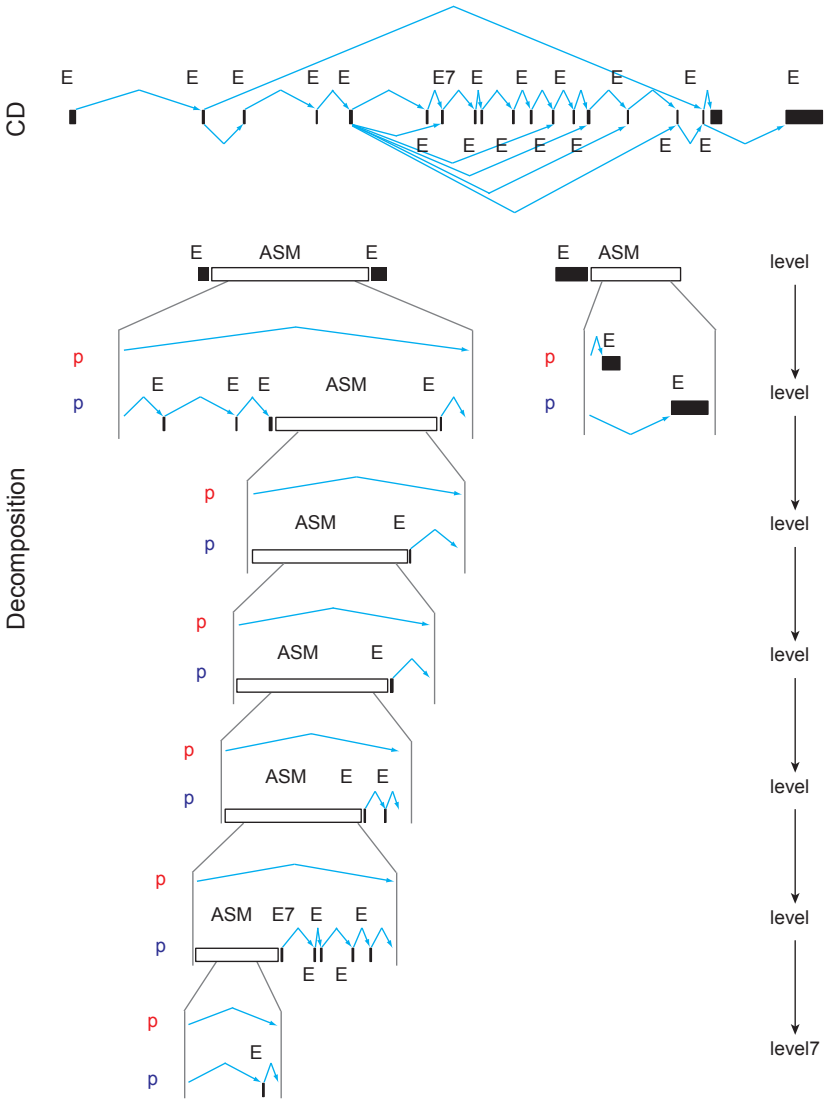


Figure 4.6: The splice graph and the ASM decomposition of gene CD44.

$E_6 - E_{13}$, $E_{14} - E_{17}$ and E_{18} (Figure 4.6). Brown *et al.* have suggested a shift in CD44 expression from variant isoforms (CD44v) to the standard isoform (CD44s) is essential in epithelial cell development and is associated with breast cancer progression. Brown et al. [2011] The alternative exons by which CD44s and CD44v differ, $E_6 - E_{13}$, are captured by three ASMs ASM_4, ASM_5 and ASM_6, where CD44s takes path p_1 in ASM_4 and CD44v takes path p_2 in all ASM_4, ASM_5 and ASM_6. Therefore, the joint

analysis of all the three ASMs will be essential for the study of the isoform transition in this gene.

Copyright © Yin Hu, 2013.

Chapter 5 The Quantification and Differential Analysis of Splicing Events

5.1 Introduction

Next, we estimate the number of transcript copies that flow through each splice path in the ASM for each individual sample. Specifically, for every ASM, we estimate the relative proportion as well as the expression level of its alternative paths in each sample. Typical Poisson-based methods such as Jiang and Wong [2009a], Srivastava and Chen [2010a] collect the number of reads falling on each exon as observations. Because only the starting position of each read contributes to the observed counts, these methods ignore the information encoded in the rest of the nucleotides such as the coverage of splice junction. The counting approach makes it infeasible to incorporate spliced reads in the model for better estimation. DiffSplice proposes a generalized model that takes into account the observed support on splice junctions in addition to exon expression to estimate the abundance of alternative paths. Such consideration is crucial for estimating alternative transcription paths since alternative splice junctions differentiate the isoforms.

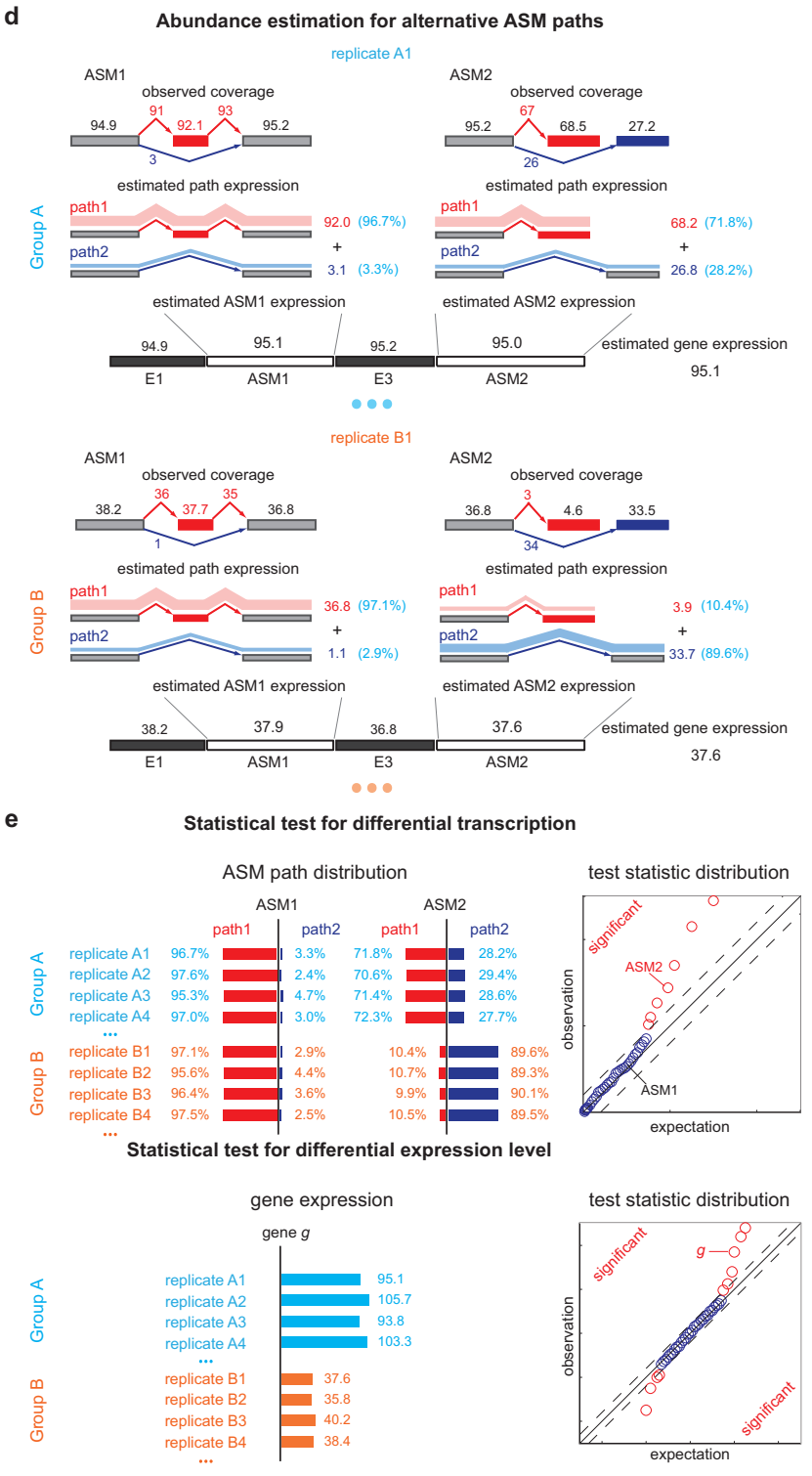


Figure 5.1: DiffSplice discovers genome-wide differential splicing events using RNA-seq data. **(d)** The abundance of the alternative transcription paths in every module is estimated, as well as the expression of the gene. **(e)** The statistical tests lastly select modules and genes with significant differences in transcription and gene expression, respectively.

5.2 Related work

5.2.1 Transcript abundance estimation

Through transcriptome assembly, *de novo* or *ab initio*, the diversity of transcripts in the sample should be revealed. Alternatively, transcripts from an annotation database such as Ensembl [Ens] and Refseq [Ref] may also be used if detecting novel isoforms is not a major concern. Then the next question is about the abundance of these transcripts. The expression level of each transcript precisely profiles the transcription of the cell and may provide insights in protein expression. Accurate measurement of transcript abundance in a given sample will enable the detection of differential expression of alternative transcripts under different conditions.

Existing transcript quantification methods can be roughly summarized into two strategies at different analyzing units: the read-centric strategy and the exon-centric strategy. In essence, the read-centric approaches set up a probabilistic model that relates the observed read alignment to the latent transcript sampling probability. The origin of a read is considered in a "fuzzy" manner, not specifying a single parent transcript for the read but characterizing with a (discrete) distribution over all possible parent transcript. The probability of observing a given alignment is then factorized into the summarization of the products of the probability that the read actually originates from the parent transcript and the conditional probability that evaluates the quality of the alignment. The latent transcript sampling probabilities are estimated by maximizing the joint likelihood of all reads, typically using an expectation-maximization scheme. Some representative methods taking this strategy

include Cufflinks [Trapnell et al., 2010], IsoEM [Nicolae et al., 2011] and RSEM [Li and Dewey, 2011].

The exon-centric approaches, on the other hand, take the expression of exons as basic observations. Taking the transcript-exon composition matrix, the expected read abundance on each exon is the cumulative abundance of all transcript isoforms that contain the exon. The transcript abundances are then solved by minimizing the overall distance from the observed read abundance on all exons in the gene to the expectations. Some representative methods include the Poisson-based models [Jiang and Wong, 2009b, Richard et al., 2010, Srivastava and Chen, 2010b] and linear regression approaches such as rQuant [Bohnert and Rätsch, 2010], IsoLasso [Li et al., 2011a] and SLIDE [Li et al., 2011a]. In MultiSplice [Huang et al., 2012], the linear model further includes a set of highly discriminative features that span multiple exons. These features utilize full read information to improve the accuracy and the identifiability of the model.

The primary challenge for transcript abundance estimation is the significant overlaps among transcripts in a gene, making it difficult to determine the original transcript of a short read unambiguously. There may exist no unique solution to the quantification problem using short read alignment, known as the *identifiability* issue [Lacroix et al., 2008, Hiller et al., 2009]. The estimation procedures may also be confounded by nonuniform read distribution. The accuracy of the abundance estimates may be altered by the break of the random sampling assumption, due to various types of sampling biases such as position-specific biases Bohnert and Rätsch [2010], Li et al. [2010a], Roberts et al. [2011], Wu et al. [2011b] and the sequence-

specific biases Li et al. [2010b], Roberts et al. [2011], Turro et al. [2011]. The effect of the biases may be computationally alleviated by evaluating the sampling probability at different positions of transcripts and the probability of observing different nucleotides (combinations of A, C, G and T) at ends of reads [Roberts et al., 2011, Huang et al., 2012]. If estimating the transcript abundance on the basis of a reference transcriptome, it is biologically unlikely that all annotated transcripts will be present in a sample. Transcripts estimated with very low abundance may be spurious and may deviate the correctness of estimates for other transcripts. In practice, the L1 regularization (known as LASSO [Cai et al., 2010, He and Lin, 2010]) is often employed to reinforce the shrinkage of the expressed transcript set [Li et al., 2011a, Huang et al., 2012].

The resulting transcript abundances are usually measured in the unit of FPKM (fragments per kilobase of transcript per million mapped reads, used in Cufflinks [Trapnell et al., 2010] and IsoLasso [Li et al., 2011a]) or averaged read coverage (used in MultiSplice [Huang et al., 2012]). If comparing transcript abundance across different samples, the abundance should be further normalized for correct assessment of differential expression [Bolstad et al., 2003, Bullard et al., 2010, Anders and Huber, 2010, Robinson et al., 2010, Dillies et al., 2012].

5.2.2 Group-wise differential test

5.3 Estimating the abundance of alternative splicing variants

5.3.1 Preliminaries.

The notations used in the abundance estimation procedure are summarized in Table 5.1. Given a transcript t and the reads from one sample, let c_i^t be the number of reads covering the ith nucleotide in t. We define the read coverage on t as the averaged number of reads covering each base in the transcript, $C_t = \frac{1}{l_t}\sum_{i=1}^{l_t} c_i^t$, where l_t denotes the exonic length of t. Then C_t is an estimator for the number of transcript copies in the sample, which provides a direct measure for the expression level of the transcript t. Similarly we define the read coverage on an exonic segment e with exonic length of l_e as $C_e = \frac{1}{l_e}\sum_{i=1}^{l_e} c_i^e$, and we use C_j to denote the number of spliced reads that pass a splice junction j. The read coverage C_e provides an estimator for the number of transcript copies that flow through the exonic segment e. The number of spliced read alignments C_j constitutes an estimator for the number of transcript copies that pass from the donor exon to the acceptor exon connected by the junction j. Therefore, we calculate the observed read coverage for every exon and the observed number of spliced read for every junction and derive estimator for transcript coverage based on the observations.

5.3.2 The normal model for the observed read coverage.

We now demonstrate a model where read coverage will be used as the observed variables for abundance estimation. Assume the sequencing procedure as a random sampling process, in which every read is sampled independently and uniformly from

Table 5.1: Notations in the abundance estimation for alternative ASM paths.

Symbol	Meaning
r	the length of a read
t	an alternative transcription path
e	an exonic segment
Δ	an ASM
l_t	the exonic length of t
l_e	the exonic length of e
N_t	the number of reads from path t
$N_{e\|t}$	the number of reads on e from t
C_e	the read coverage exonic segment e
$C_{e\|t}$	the read coverage on exonic segment e from transcript t
$A_{t,e}$	a Boolean variable (1 or 0) indicating whether t includes e
m	the total number of exonic segments and splice junctions in Δ
n	the number of alternative transcription paths in Δ
N	the total number of reads in Δ
$\mathbf{q}$	the relative proportion of the alternative paths in Δ
Γ	the estimated expression for an alternative path or an ASM

every possible nucleotide in the transcripts Jiang and Wong [2009a]. For a single transcript t in an ASM, the probability that a read from t falls in e is $p_{e|t} = \frac{l_e}{l_t}$. Given N_t, the total number of reads from t, the number of reads falling in segment e $N_{e|t}$ follows a binomial distribution with parameters N_t and $p_{e|t}$, $N_{e|t} \sim Bin(N_t, p_{e|t})$. When N_t is sufficiently large, the binomial distribution can be well approximated using a normal distribution with mean $N_t p_{e|t}$ and variance $N_t p_{e|t}(1 - p_{e|t})$, written as $N_{e|t} \dot{\sim} N(N_t p_{e|t}, N_t p_{e|t}(1 - p_{e|t}))$. Let r denote the length of a read. The value of $\frac{N_{e|t} r}{l_e}$ represents the read coverage on e contributed by t, $C_{e|t}$, whereas the value of $\frac{N_t r}{l_t}$ represents the read coverage on t, C_t. Therefore, we have $\frac{N_{e|t} r}{l_e} \dot{\sim} N(\frac{N_t p_{e|t} r}{l_e}, \frac{r^2}{l_e^2} N_t p_{e|t}(1 - p_{e|t}))$, equivalently

$$C_{e|t} \dot{\sim} N(C_t, \frac{r(l_t - l_e)C_t}{l_t l_e}). \tag{5.1}$$

For a splice junction j, its length l_j is defined to equal the read length r, which is the length of the exonic region where reads start in this region can cover the splice junction. The number of spliced reads from t that covers j, $C_{j|t}$, still follows the normal distribution in Equation 5.1.

From Equation 5.1, $C_{e|t}$ and $C_{j|t}$ are unbiased for C_t. The variance of $C_{e|t}$ varies according to the coverage C_t and the segment length l_e. Dividing the difference between $C_{e|t}$ and C_t by C_t, we have the ratio $\frac{C_{e|t}-C_t}{C_t}$ following a normal distribution $N(0, \frac{r(l_t-l_e)}{C_t l_t l_e})$. Higher coverage and longer segments lead to estimators with smaller variance of the relative deviation from the true transcript coverage, which we demonstrate in the simulated results (Figure 5.3).

5.3.3 Estimation of alternative ASM path abundance

Consider an ASM Δ with totally m exonic segments and splice junctions. Assume Δ consists of n alternative transcription paths. The exonic length of a path t is hence given as $l_t = \sum_{i=1}^{m} A_{t,i} l_i$, where $A_{t,i} = 1$ if path t covers the i-th exonic segment and $A_{t,i} = 0$ otherwise. Let $\mathbf{q} = \{q_1, q_2, \cdots, q_n\}$ denote the relative proportions of the alternative paths, with $\sum_{i=1}^{n} q_i = 1$. The probability of a read falling into path t is then written as $p_t = \frac{q_t l_t}{\sum_{i=1}^{n} q_i l_i}$, with $\sum_{i=1}^{n} p_i = 1$. Assume the number of reads sampled from Δ follows a Poisson distribution with parameter N, where N represents the expression of Δ in the sample accounting for the depth of sequencing and the length of Δ Jiang and Wong [2009a]. The number of reads sampled from path t, N_t,

then follows a Poisson distribution with parameter $N \cdot p_t$, *i.e.*,

$$N_t \sim Poisson(N \cdot p_t). \tag{5.2}$$

We hence derive the maximum likelihood estimators for the path proportion, $\mathbf{q}$, and the expected total number of reads in Δ, N. With observed read coverage $C_1, C_2, \cdots, C_m$ on every exonic segment and splice junction, the likelihood of $\mathbf{q}$ and N is the joint density of C_1 through C_m under $\mathbf{q}$ and N,

$$L(\mathbf{q}, N|data) = L(\mathbf{q}, N|C_1, \cdots, C_m) = P(C_1, \cdots, C_m|\mathbf{q}, N)$$

We assume that $C_1, C_2, \cdots, C_m$ are mutually independent. The likelihood function can be factorized as

$$L(\mathbf{q}, N|C_1, \cdots, C_m) = \prod_{t=1}^{n} \prod_{i=1}^{m} f(C_{i|t}|N_t) g(N_t), \tag{5.3}$$

where $f(\cdot)$ is the density of the exonic/junction coverage distribution in Equation 5.1 and $g(\cdot)$ is the density of the transcript read count distribution in Equation 5.2.

5.3.4 Abundance estimation in ASM

Consider an ASM with n alternative transcription paths and m features (exonic segments and splice junctions). We define $A_{t,e}$ as an indicator for the presence of a feature e in transcription path t, with value of 1 if t covers e and 0 otherwise. The indicators for the presence of every exon/junction in each path form an $n \times m$ indicator matrix A.

Derivation of likelihood function

Let $C_{e|t}$ denote the coverage on the eth feature from the tth path. Under the independence assumption, the likelihood can be factorized as

$$\begin{aligned} L(\mathbf{q}, N|C_1, \cdots, C_m) &= P(C_{1|1}, \cdots, C_{1|n}, C_{2|1}, \cdots, C_{2|n}, \cdots, C_{m|1}, \cdots, C_{m|n}|\mathbf{q}, N) \\ &= \prod_{t=1}^{n} P(C_{1|t}, C_{2|t}, \cdots, C_{m|t}) \\ &= \prod_{t=1}^{n} P(C_{1|t}, C_{2|t}, \cdots, C_{m|t}|N_t)P(N_t) \\ &= \prod_{t=1}^{n}\prod_{i=1}^{m} P(C_{i|t}|N_t)P(N_t) \\ &= \prod_{t=1}^{n}\prod_{i=1}^{m} f(C_{i|t}|N_t)g(N_t), \end{aligned}$$

where $f(\cdot)$ is the density of $N(C_t, \frac{r(l_t-l_e)C_t}{l_t l_e})$ and $g(\cdot)$ is the density of $Poisson(\lambda_t)$, $\lambda_t = N \cdot p_t$.

Maximum likelihood estimators

The maximum likelihood estimator for $\mathbf{q}$ and N are the ones that maximize the likelihood,

$$(\hat{\mathbf{q}}, \hat{N}) = \arg\max_{\mathbf{q},N} L(\mathbf{q}, N|data).$$

$$
\begin{aligned}
& l(\mathbf{q}, N|C_1, \cdots, C_m) \\
= \; & \log L(\mathbf{q}, N|C_1, \cdots, C_m) \\
= \; & \sum_{t=1}^{n}[\log(g(N_t)) + \sum_{i=1}^{m} \log f(C_{i|t}|N_t)] \\
= \; & \sum_{t=1}^{n}\{\log \frac{e^{-\lambda_t}\lambda_t^{N_t}}{N_t!} + \sum_{i=1}^{m} \log[\frac{1}{\sqrt{2\pi r(l_t - l_i)C_t/(l_t l_i)}} e^{-\frac{(C_{i|t}-C_t)^2}{2r(l_t-l_e)C_t/(l_t l_e)}}]\} \\
= \; & \sum_{t=1}^{n}\{-\lambda_t + N_t \log \lambda_t - \log N_t! + \sum_{i=1}^{m}[\frac{1}{2}\log l_t + \frac{1}{2}\log l_i - \frac{1}{2}\log 2\pi - \frac{1}{2}\log r \\
& -\frac{1}{2}\log(l_t - l_i) - \frac{1}{2}\log C_t - \frac{(C_{i|t} - C_t)^2}{2r(l_t - l_e)C_t/(l_t l_e)}]\}
\end{aligned}
$$

EM algorithm for deriving estimators

We then use the expectation maximization (EM) algorithm to derive the maximum likelihood estimators for $\mathbf{q}$ and N (Supplementary Section 2). In addition to estimating transcription path proportions, the EM algorithm also calculates the expected expression of each transcription path, $\Gamma_1, \Gamma_2, \cdots \Gamma_n$. Then the expected expression of Δ sums up the expected expression of all transcription paths in Δ, $\Gamma_\Delta = \sum_{t=1}^{n} \Gamma_t$, forming an estimator for the total number of transcript copies passing through Δ.

The expectation maximization (EM) algorithm to find the maximum likelihood estimator $\hat{\mathbf{q}}$ and $\hat{N}$ is detailed as the following.

1. *E-step*:

Denoting the values of q_t at step v as $q_t^{(v)}$, we first calculate the conditional expectation of C_t conditioning on $q_t^{(v)}$. Let $C_{(1)}, C_{(2)}, \cdots, C_{(m')}$ be the read coverage of the exonic segments that are in path t, *i.e.*, $A_{t,e} = 1$ if $e \in \{(1), (2), \cdots, (m')\}$ and

$A_{t,e} = 0$ otherwise. Let $\hat{C_{e|t}}$ denote the expected coverage on exonic segment e from t, $\hat{C_{e|t}} = \frac{p_e q_t^{(v)} A_{t,e}}{\sum_{j=1}^{n} p_e q_j^{(v)} A_{j,e}} \cdot C_e$. Let $k_{t,e}$ denote $\frac{r(l_t - l_e)}{l_t l_e}$, so we have $C_{e|t} \sim N(C_t, k_{t,e} C_t)$. Therefore, the conditional expectation of C_t is the maximum likelihood estimator that maximizes the joint density of the m' normal densities,

$$\hat{C_t} = E_{q_t^{(v)}}\left[C_t | C_{(1)}, C_{(2)}, \cdots, C_{(m')}\right] = \frac{-m' + \sqrt{m'^2 + 4\sum_{i=1}^{m'} k_{t,(i)}^{-1} \sum_{i=1}^{m'} k_{t,(i)}^{-1} \hat{C_{(i)|t}}^2}}{2\sum_{i=1}^{m'} k_{t,(i)}^{-1}}$$

The expected number of reads on path t is hence calculated as $\hat{N_t} = \frac{\hat{C_t} l_t}{r}$.

2. *M-step:* Then we derive the parameters that maximize the conditional likelihood on $\hat{N_t}$:

$$\begin{aligned} & \text{Set } \frac{\partial L}{\partial N} \text{ to } 0 \\ \Rightarrow \quad & -1 + \sum_{t=1}^{n} N_t \frac{1}{\hat{N}} = 0 \\ \Rightarrow \quad & \hat{N} = \sum_{t=1}^{n} \hat{N_t} \end{aligned}$$

$$\begin{aligned} & \text{Set } \frac{\partial L}{\partial q_t} \text{ to } 0 \\ \Rightarrow \quad & \sum_{t=1}^{n} \left(-\frac{d\lambda_t}{d\hat{q}_t} + N_t \cdot \frac{1}{\lambda} \cdot \frac{d\lambda_t}{d\hat{q}_t} \right) = 0 \\ \Rightarrow \quad & \sum_{t=1}^{n} \left((\frac{N_t}{N} - 1) \cdot \frac{d\lambda_t}{d\hat{q}_t} \right) = 0 \\ \Rightarrow \quad & \hat{q}_t^{(v)} = \frac{\hat{N_t} \cdot \sum_{j=1, j\neq t}^{n} \left(\hat{q}_j^{(v-1)} (\sum_{i=1}^{m} p_i A_{j,i}) \right)}{(\hat{N} - \hat{N_t}) \cdot (\sum_{i=1}^{m} p_i A_{t,i})} \end{aligned}$$

5.3.5 Estimation of gene expression

Within a gene G, the abundance estimation procedure starts from the minimal ASMs, *i.e.*, the ASMs in the bottom level of the decomposition hierarchy, then propagates

towards the top of the hierarchy. During inference within an ASM Δ, all ASMs nested in Δ must have performed the alternative path abundance estimation and hence are treated as single exonic segments, using their estimated expression as the exonic coverage. The estimator for the expression of gene G, Γ_G, is hence the mean expression of all the exonic segments and ASMs that directly constitute G (or in the decomposition hierarchy all the children of G on the first level). This estimator provides a direct measure for the expected total number of transcript copies in gene G in the RNA-seq sample.

Copyright © Yin Hu, 2013.

5.4 Statistical test for differential transcription

Differential expression under different conditions may exhibit in two aspects. At the gene level, the difference in a gene's expression level measures the change of the total expression of all the transcripts in this gene (*differential gene expression level*). At the transcript level, the difference in the relative proportion of alternative transcription paths reflects the regulation on the expression of individual transcripts (*differential gene transcription*). In many cases, these two types of differentiations positively correlate, because the overall expression level of a gene is made up additively of the expression levels of all the transcripts that the gene code. The up/down-regulation of one or more transcripts may result in the up/down-regulation of the entire gene. However, the transcript-level differential expression analysis, the differential transcription analysis, may answer two additional questions of high importance — which subset of the transcripts in the genes have been regulated from one condition to another, and whether the on/off of one subset of transcripts associates with the off/on of another subset of transcripts in the gene. While the former provides much higher accuracy and a much higher resolution about expression regulation, which may directly point to prominent protein isoforms, the latter may reveal the phenomenon of "isoform switching", which may help understand the regulation network.

Therefore, we have developed separate statistical test procedures for the detection of differential expression at these two levels. In this chapter, we focus on the group-wise statistical analyses, in which the grouping of the samples is pre-assumed and the aim is to test the change of gene/transcript expression from one group to another.

Let $S_1, S_2, \cdots, S_k$ denote the k groups of samples in the dataset. For example, a dataset with 2 sample groups may compare a diseased group to a normal control group, and a dataset with more than 2 sample groups may compare among certain tumor samples classified as different subtypes. For a sample group S_i, let n_i be the size of the group, the number of samples classified as group i. Let $S_i = \{s_i^1, s_i^2, \cdots, s_i^{n_i}\}$ denote set of samples in S_i.

5.4.1 Two-group differential transcription

We first consider the scenario of two-group comparison, that is, $k = 2$.

Transcription profile characterized by isoform proportion. Within every sample, the transcription is profiled at every ASM, the loci where the transcripts of a gene diverge. Let $Q_1 = \{\mathbf{q}_1^1, \mathbf{q}_1^2, \cdots, \mathbf{q}_1^{n_1}\}$ and $Q_2 = \{\mathbf{q}_2^1, \mathbf{q}_2^2, \cdots, \mathbf{q}_2^{n_2}\}$ denote the estimated path proportion of an ASM Δ in each sample. The path proportion distribution $\mathbf{q}_i^j$ then describes the transcription profile of Δ in sample s_i^j. Every distribution $\mathbf{q}_i^j$ is a t-dimensional real vector with t being the number of alternative transcription paths in Δ. Every dimension of $\mathbf{q}_i^j$ should lie in $[0, 1]$, and sum of all dimensions should equal 1, $\|\mathbf{q}_i^j\|_1 = 1$.

Let $\bar{\mathbf{q}_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{q}_1^i$ and $\bar{\mathbf{q}_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{q}_2^i$ denote the mean distributions of the two sample groups. The hypotheses of the two-group differential transcription test are then

Null the mean path distributions of the two groups are the same, $\bar{\mathbf{q}_1} = \bar{\mathbf{q}_2}$;

Alternative the mean path distributions of the two groups are not the same, $\bar{\mathbf{q}_1} \neq$

$\bar{\mathbf{q}}_2$.

Jensen-Shannon divergence. We us the Jensen-Shannon divergence (JSD) Lin [1991] to quantify the dissimilarity between two distributions as a real value between 0 and 1. Let $\mathbf{p} = (p_1, \cdots, p_t)^T$ and $\mathbf{q} = (q_1, \cdots, q_t)^T$ be two t-dimensional distributions. The Jensen-Shannon divergence (JSD) is calculated as

$$JSD(\mathbf{p}||\mathbf{q}) \quad = \quad \frac{1}{2}\left(KLD(\mathbf{p}||\mu) + KLD(\mathbf{q}||\mu)\right), \tag{5.4}$$

where $\mu = \frac{1}{2}(\mathbf{p} + \mathbf{q})$ is the mean distribution of $\mathbf{p}$ and $\mathbf{q}$, and KLD is the Kullback-Leibler divergence (Kullback and Leibler [1951]) defined as

$$KLD(\mathbf{p}||\mathbf{q}) \quad = \quad \sum_{j=1}^{t} p_j \log \frac{p_j}{q_j}. \tag{5.5}$$

The difference of transcription between two samples s_1^i and s_2^j is then measured by the metric the square root of the Jensen-Shannon divergence, $\sqrt{JSD(\mathbf{q}_1^i||\mathbf{q}_1^j)}$.

Test statistic for differential gene transcription.

To select significant differences in transcription, we look for ASMs with significant difference in path distributions between the two groups but consistent path distributions within each group. We define the between-group difference as the divergence between the group mean distributions,

$$x_\Delta = \sqrt{JSD(\bar{\mathbf{q}}_1||\bar{\mathbf{q}}_2)}. \tag{5.6}$$

The within-group variance of each group is defined as

$$s_\Delta = \sqrt{c[\sum_{j=1}^{n_1} JSD(\mathbf{q}_1^j||\bar{\mathbf{q}}_1) + \sum_{j=1}^{n_2} JSD(\mathbf{q}_2^j||\bar{\mathbf{q}}_2)]}, \tag{5.7}$$

where $c = \frac{n_1+n_2}{n_1 n_2 (n_1+n_2-2)}$ is the normalization constant.

Abundance estimation on ASMs with low expression often associates with higher instability. Therefore, we add σ_Δ as a penalty for low expression, based on a logistic function of the averaged estimated expression of the ASM Γ_Δ,

$$\sigma_\Delta = (2 - \frac{2}{1 + e^{-\phi \Gamma_\Delta}}) \cdot s_{max}, \text{ for } \Gamma_\Delta \geq 0, \tag{5.8}$$

where ϕ adjusts the penalized expression range of low ASM expression (*e.g.*, $\phi = 1$ for penalizing ASMs with estimated expression less than around 6 while assigning negligible penalty to ASMs with higher expression) and s_{max} denotes the largest variance among all ASMs in the data.

Therefore, the relative difference in transcription of an ASM Δ is in the form

$$d_\Delta = \frac{x_\Delta}{s_\Delta + \sigma_\Delta}, \tag{5.9}$$

measuring the extent how the distributions over alternative paths within the ASM consistently differ between the two groups.

Permutation test. An empirical distribution of relative difference can be obtained by calculating test statistics after permuting samples across groups Tusher et al. [2001]. Suppose totally M ASMs are tested for differential transcription. The relative transcriptional difference is calculated for every ASM, and the order statistics are collected, $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(M)}$. Under each permutation p, order statistics of relative differences could also be calculated in the same way: $d^{*p}_{(1)} \leq d^{*p}_{(2)} \leq \cdots \leq d^{*p}_{(M)}$. Averaging order statistics from all permutations, we have the expected relative difference in transcription: $\bar{d^*_{(i)}} = \frac{1}{|P|} \sum_{p \in P} d^{*p}_{(i)}$ for $1 \leq i \leq M$, where P is the set of all permutations and $|P|$ is the number of permutations.

Statistical significance. Significant changes on transcription are concluded based on the extent of disagreement between calculated and expected test statistics. Given a threshold δ_{trans}, an ASM with relative transcription difference of $d_{(i)}$ is accepted to have significant difference on transcription, if $|d_{(i)} - \bar{d}^*_{(i)}| > \delta_{trans}$. The choice of δ_{trans} is monitored by its associated false discovery rate, which we define next.

False discovery rate (FDR). At a cutoff of δ_{trans}, the quantity of falsely discovered ASMs in each permutation is estimated as the number of ASMs such that $|d^{*p}_{(i)} - \bar{d}^*_{(i)}| > \delta_{trans}$. The FDR for differential transcription is hence estimated as the averaged number of falsely discovered ASMs over all permutations, divided by the total number of ASMs.

5.4.2 Differential transcription among more than two groups

Next we discuss the differential transcription test in a dataset with more than 2 sample groups, that is, $k > 2$.

Let $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \cdots, \bar{\mathbf{q}}_k$ denote the mean distributions of the k sample groups. Let $\bar{\mathbf{q}}_\cdot$ denote the grand mean distribution, the averaged mean distribution over all samples in all groups,

$$\bar{\mathbf{q}}_\cdot = \frac{\sum_{j=1}^{k} n_j \cdot \bar{\mathbf{q}}_j}{\sum_{j=1}^{k} n_j}. \tag{5.10}$$

The hypotheses being tested for the k-group differential transcription are then

Null the mean path distributions of the k groups are all the same, $\bar{\mathbf{q}}_\cdot = \bar{\mathbf{q}}_1 = \bar{\mathbf{q}}_2 = \cdots = \bar{\mathbf{q}}_k$;

Alternative there exist at least two groups whose mean path distributions are not the same, $\exists i \neq j, \bar{\mathbf{q}}_i \neq \bar{\mathbf{q}}_j$.

The test statistic of differential transcription on ASM Δ for $k > 2$ groups is changed to

$$x_\Delta = \frac{\sum_{j=1}^{k} n_j \cdot JSD(\bar{\mathbf{q}}_j || \bar{\mathbf{q}}_\cdot)}{k-1} \tag{5.11}$$

and

$$s_\Delta = \frac{\sum_{j=1}^{k} \sum_{h=1}^{n_j} JSD(\mathbf{q}_j^h || \bar{\mathbf{q}}_j)}{\sum_{j=1}^{k} n_j - k}. \tag{5.12}$$

The relative difference in transcription of the ASM Δ is still measured by the ratio of the difference among group means x_Δ against the within-group variance s_Δ

$$d_\Delta = \frac{x_\Delta}{s_\Delta + \sigma_\Delta}. \tag{5.13}$$

5.4.3 Differential gene expression

Based on the estimators for gene expression level derived in the previous section, we use the same method as proposed in SAM Tusher et al. [2001] to test for difference in gene expression under different conditions (groups).

The relative difference in expression level of a gene i is written as $d_i = \frac{r_i}{s_i + s_0}$, which measures the ratio of difference among group means over within-group variances. A normalization term s_0 learned from data Tusher et al. [2001] is added to the denominator in case the variances are too small. Because genes with low expression often have small variance, this normalization term also asks for larger gap on group means from lowly expressed genes to evidence significant difference. Comparing expression

level of gene i in two groups of samples, let x_{11}, x_{12}, $\cdots$, x_{1n_1} be estimated gene coverage of gene i in group 1 and x_{21}, x_{22}, $\cdots$, x_{2n_2} be estimated gene coverage in group 2. The difference between the two group means is written as

$$r_i = \bar{x_{2\cdot}} - \bar{x_{1\cdot}};$$

the normalized mean of within-group variance is written as

$$s_i = \sqrt{c[\sum_{j=1}^{n_1}(x_{1j} - \bar{x_{1\cdot}})^2 + \sum_{j}^{n_2}(x_{2j} - \bar{x_{2\cdot}})^2]},$$

where $c = \frac{n_1+n_2}{n_1 n_2(n_1+n_2-2)}$ is the normalization constant.

5.5 Simulation studies of group-wise differential transcription analysis

The following set of experiments first evaluated the accuracy of DiffSplice on datasets simulated on the entire human transcriptome with varying sampling depth and varying degrees of 5' or 3' positional bias. We then compared DiffSplice with the state of the art methods including Cufflinks and Flow Difference Metric (FDM) on the simulated dataset used by Singh *et al.* Singh et al. [2011].

5.5.1 Simulation of RNA-seq datasets.

We developed an in-house simulator to generate two RNA-seq datasets on human transcriptome. In each dataset, we generated pairs of RNA-seq samples under various sampling depth or sampling bias. For every sample, the simulator randomly generates relative expression profiles for the transcripts, based on the user-provided human

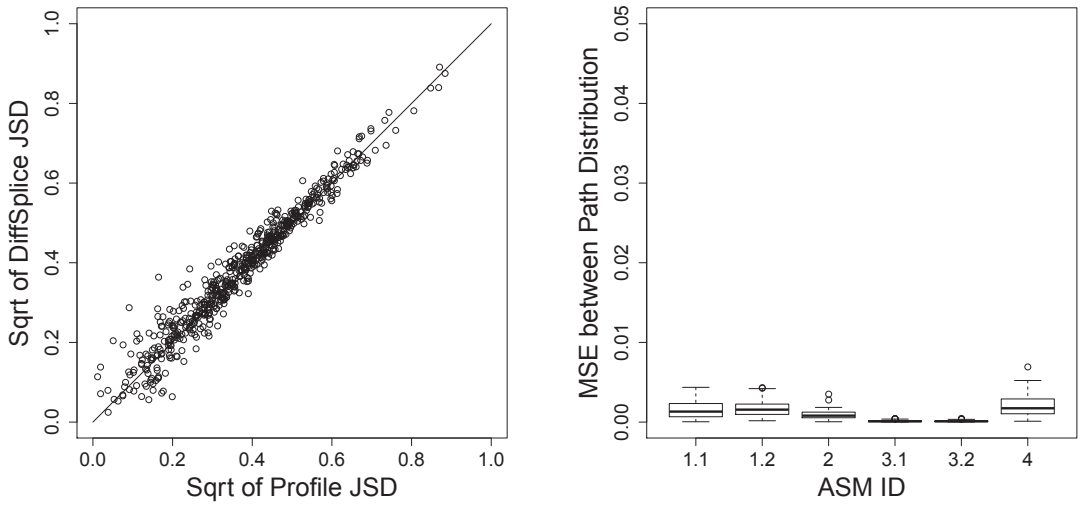
transcriptome annotation. A number of cDNA molecules are then assigned to every transcript according to its expression level and the size of the dataset. A cDNA library is hence constructed through steps as amplification and size-selection, and RNA-seq reads are sampled from the cDNA library.

For every pair of samples, we first calculate their transcriptional difference at each ASM based on the transcript annotation and expression profiles used in generating the RNA-seq data, referred to as the *profile JSD*. The difference in ASM estimated by DiffSplice directly from the RNA-seq reads is referred to as the *DiffSplice JSD*. The profile JSD reflects the ground truth difference in each ASM, while the DiffSplice JSD is an estimation from sampled reads. We calculate the Pearson correlation between the two as a measure for the accuracy of the estimated difference, denoted by the *JSD correlation* We also consider a complementary measure for every ASM, the mean squared error (MSE), which calculates the error of the estimated path distribution from the distribution in the expression profile. We average the MSE from both samples in a pair-wise comparison and denote as the *MSE of path distribution*.

5.5.2 Analyzing accuracy on highly complex gene model

We simulated 100 runs of experiments on this gene. In each run, 2 sets of RNA-seq reads were generated by 2 independently created transcript expression profiles. Every set of reads had 50K 50bp single-end reads. In Figure 5.2a, every single dot represents an ASM in one run. All ASMs have the divergence estimated by DiffSplice very close to the profile divergence, with a Pearson correlation as high as 0.974. This precision in quantifying sample-sample divergence results from the accuracy in path abundance

estimation. Figure 5.2b plots the distribution of the MSE between path distribution for every single ASM. All 6 ASMs have the majority of their MSE below 0.005 with mean close to 0 and small variances, showing the accuracy of the abundance estimator developed in DiffSplice.



(a) (b)

Figure 5.2: Evaluation of DiffSplice on simulated dataset of gene VEGFA. (a) Comparison between difference calculated from sampling profile and difference estimated by DiffSplice, measured by the square root of JSD. The Pearson correlation is 0.974. (b) The mean squared error (MSE) between sampling profile and estimated alternative path distribution, averaged between the two samples. The abundance estimation procedure of DiffSplice has very low error on all the 6 ASMs.

5.5.3 Human transcriptome under varying sampling depth.

We first study the effect of the sampling depth on the abundance estimation. We simulated 10 pairs of samples on human transcriptome, from 10M (10%) reads to 100M (100%) reads. For each sample, 2x75bp paired-end reads with average insert-size of 100bp were generated. Genes with averaged read coverage per base greater than 10 were picked to compare the difference by profile and the difference derived by

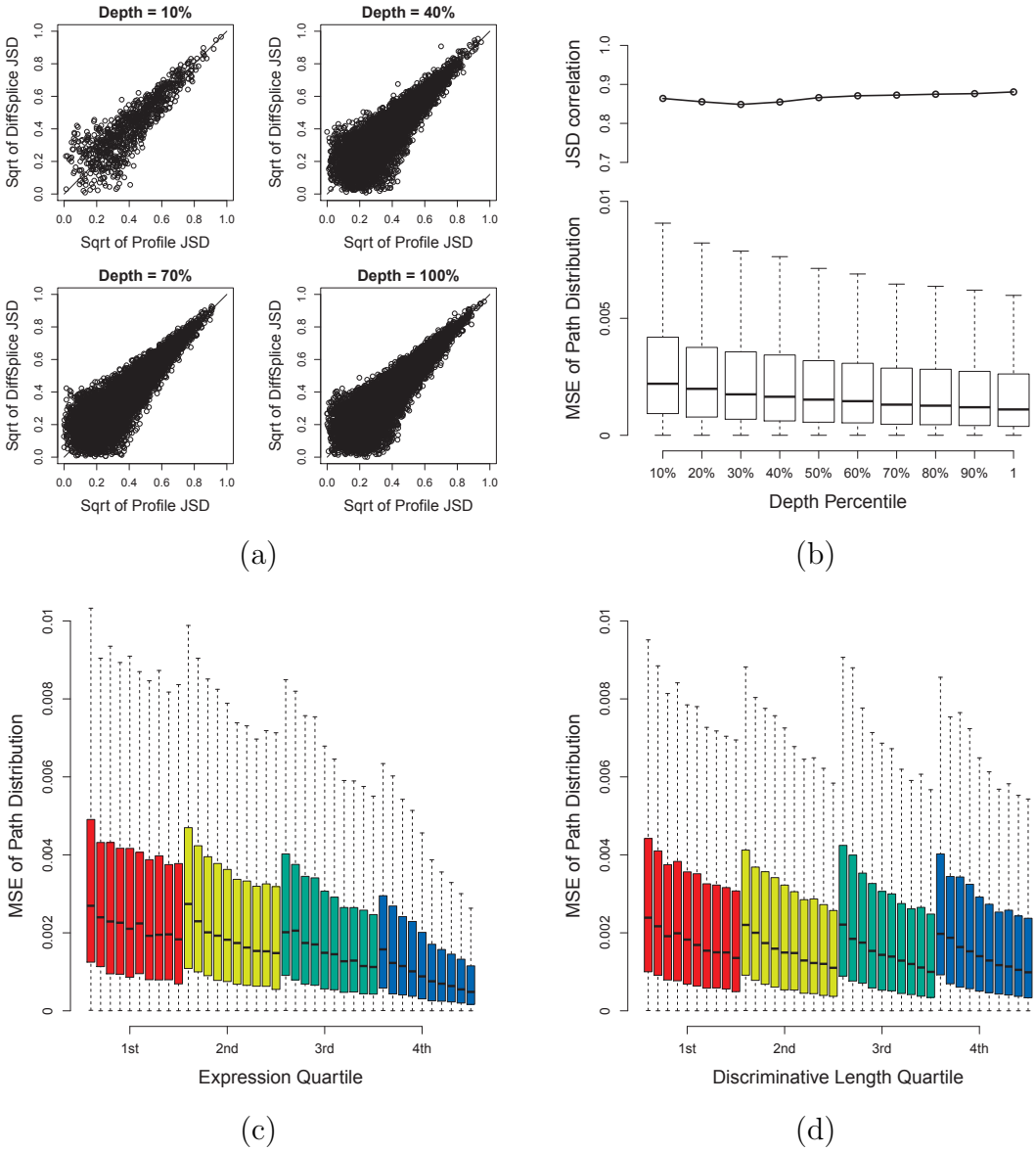


Figure 5.3: Evaluation of DiffSplice on simulated dataset under different sampling depth. **(a)** Scatterplot of profile JSD and DiffSplice JSD at different sampling depth. **(b)** JSD correlation and MSE of path distribution at different sampling depth (from 10% to 100%). **(c)** MSE of path distribution grouped by different expression quartile. **(d)** MSE of path distribution grouped by different discriminative length quartile. Within each quartile group, the box plot of the MSE is plotted for every read set (from left to right: the read sets with sampling depth percentile of 10% through 100%).

DiffSplice. Figure 5.3a shows the scatterplots of profile JSD against JSD estimated by ASM in read sets of 10M (10%), 40M (40%), 70M (70%), and 100M (100%) reads. The data with relatively lower sampling depth (*e.g.* depth = 10%) shows less points than the data with higher sampling depth because it covers less ASMs. But all sets have most points close to the diagonal, indicating minimal deviation between profile JSD and estimated JSD. The correlations range from 0.85 to 0.88 (Figure 5.3b). Higher JSD correlation is achieved by increasing the sampling depth, while the MSE of path distribution also decreases. Figure 5.3c separates all ASMs into 4 quartile groups according to their expression level and compares the distribution of MSE in each group. ASMs with higher expression separate randomness of read sampling and result in more stable estimates. As expected, the upper 2 quartiles exhibit better estimates than the lower 2 quartiles, in terms of both smaller mean and lower variance.

Besides the expression of the ASMs, the variance of the abundance estimator is also related to the *discriminative length*, the length of the exonic regions that are specific to a path in an ASM. Figure 5.3d groups all ASMs into 4 quartiles according to the discriminative length. ASMs with larger discriminative length are also expected to be more robust to random sampling errors and have higher accuracy on discriminating difference between path distributions. The lowest quartile has slightly higher MSE than the rest 75% ASMs. In contrast, the MSE sharply decreases in all groups, emphasizing the impact of sampling depth over discriminative length in improving abundance estimation accuracy.

5.5.4 Human transcriptome under varying sampling bias.

Methods that estimate transcript abundance are typically designed under the assumption that the RNA-seq fragments are sampled independently and the sampling position is uniformly distributed along the transcript from which the fragments originate. The transcript inference and thereafter the evaluation of differential expression may be altered if sampling bias is introduced by sample preparation protocols. Two types of sampling bias are commonly observed in RNA-seq data, namely, position-specific bias and sequence-specific bias Bohnert and Rätsch [2010], Srivastava and Chen [2010a], Olejniczak et al. [2010], Roberts et al. [2011].

We specifically looked at 3' bias that is a typical position-specific bias. To simulate the data, we introduce a parameter β to represent the degree of sampling bias, such that $1 + \beta$ equals to the ratio of the sampling probability at the last base in the 3' end of a transcript over the sampling probability at the first base in the 5' end of the transcript. The sampling probability at a middle bases t is then calculated as a linear interpolation, $Prob_t = Prob_{5'} \cdot (1 + \beta \cdot l_t/l)$, where l_t denotes the distance from the base t to the 5' end of the transcript and l denotes the length of the transcript.

We simulated 11 read sets on human transcriptome under β from 0 to 2.0. Figure 5.4a shows the scatterplots of the profile JSD against the DiffSplice JSD in read sets under no bias and bias of $\beta = 0.6$, 1.2, and 1.8. All sets have most estimated JSD close to profile JSD, with no significant effect of sampling bias. This is consistent with Figure 5.4b where the correlations range from 0.878 to 0.887. The MSE is slightly lower when no bias is introduced but remains roughly unchanged as β increases, in-

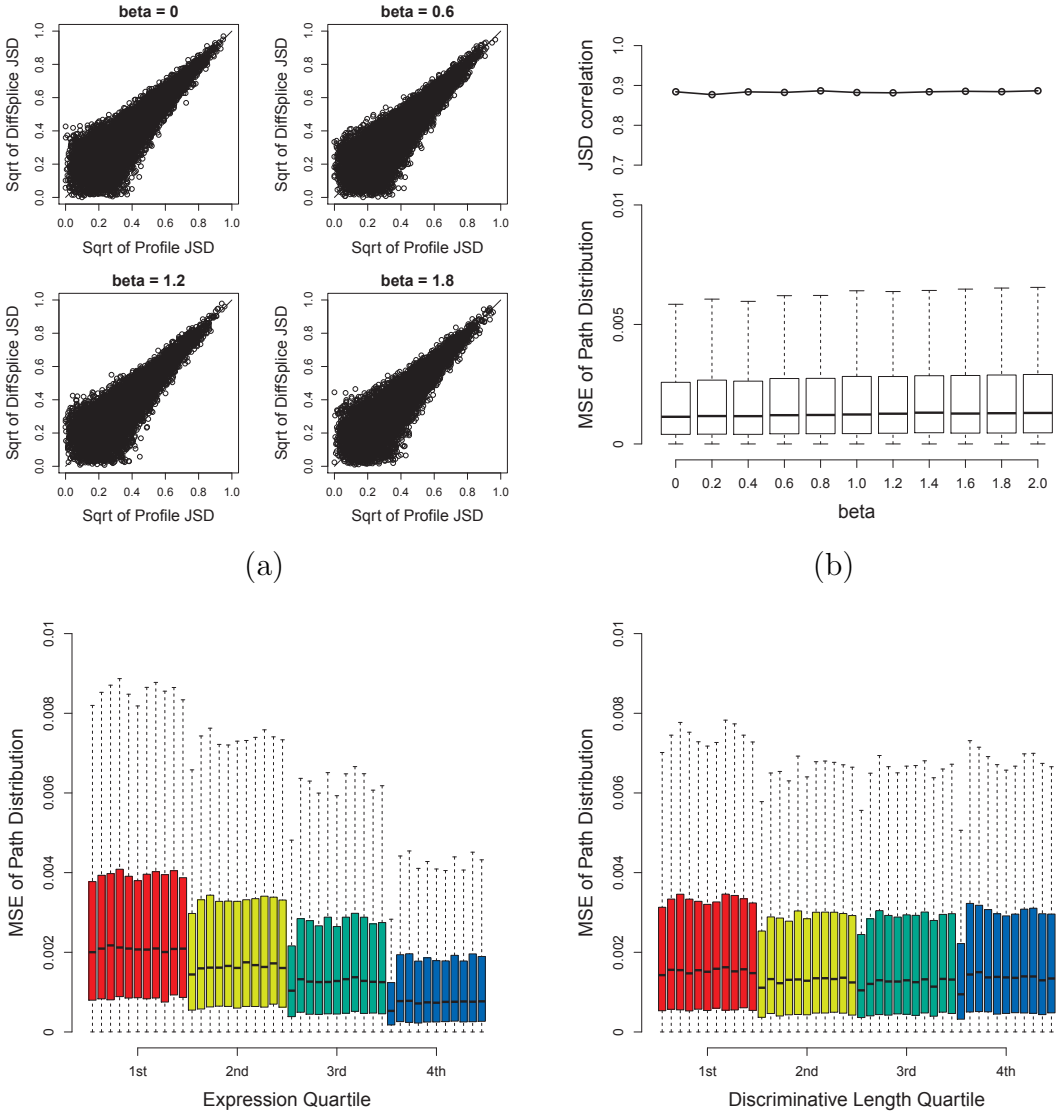


Figure 5.4: Evaluation of DiffSplice on simulated dataset in the presence of position-specific sampling bias. **(a)** Scatterplot of profile JSD and DiffSplice JSD at different β. **(b)** JSD correlation and MSE of path distribution at different β (from 0 to 2). **(c)** MSE of path distribution grouped by different expression quartile. **(d)** MSE of path distribution grouped by different discriminative length quartile. Within each quartile group, the box plot of the MSE is plotted for every read set (from left to right: the read sets with beta of 0 through 2).

dicating the robustness against altered sampling distribution of the alternative path estimation by DiffSplice. In Figure 5.4c and Figure 5.4d, ASMs are again grouped into quartile groups according to their expression level and discriminative length. While the expression level still dominates the accuracy of path abundance estimation, no significant effect of sampling bias is observed in all groups.

5.5.5 Differential transcription between two groups of samples.

We further applied our method to the two simulated datasets used in the evaluation of FDM Singh et al. [2011]. Over 2100 genes with at least two transcripts were simulated in the two tissues, each tissue having four replicates. The square root of the JSD between transcript profiles of the two tissues was calculated for each gene to suggest the "true" transcriptional difference. The coverage of each gene was calculated to measure the expression level. Genes with coverage larger than 1 were chosen for comparison. In addition to DiffSplice, three other methods (FDM, Cuffdiff with annotation, Cuffdiff without annotation) were also applied on this dataset. FDM was run using no transcriptome annotation information. With FDR less than 0.01, DiffSplice reported 887 genes with significant difference on transcription. At confidence level of 0.05, FDM, Cuffdiff with annotation, and Cuffdiff without annotation reported 722, 931, and 530 differentially transcribed genes, respectively.

Figure 5.5 plot the genes coordinated by the square root of its profile JSD and the logarithm of its coverage. The genes with significant differences on transcription identified by each method are represented by red dots. The genes with insignificant differences are represented by blue circles. Along the x-axis, the majority of the sig-

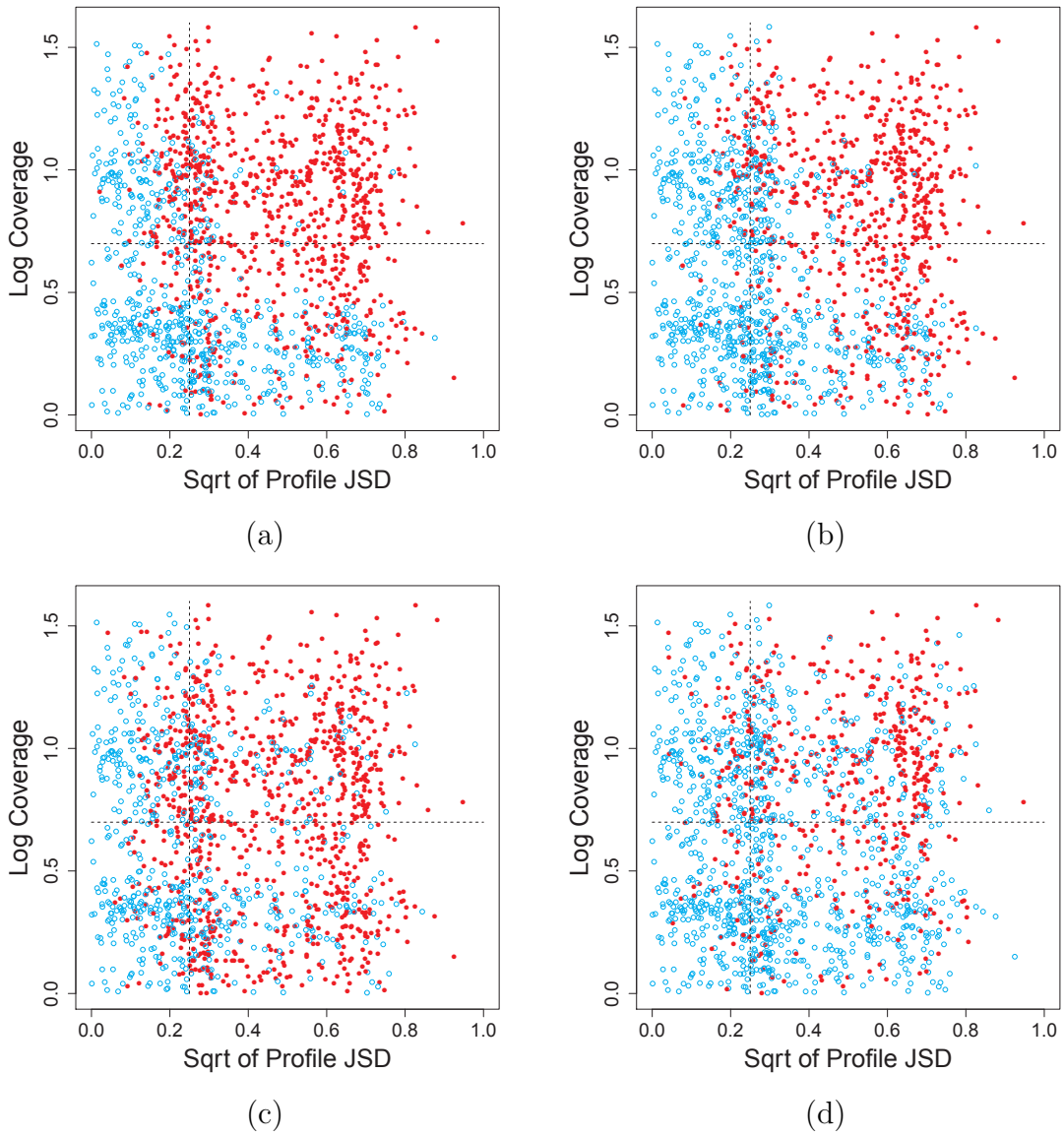


Figure 5.5: Comparison among DiffSplice, FDM, and Cufflinks on simulated dataset of human transcriptome: scatterplot of coverage against profile JSD for results of (a) DiffSplice, (b) FDM, (c) Cufflinks with annotation, and (d) Cufflinks without annotation, respectively. The majority of the differentially transcribed genes identified by DiffSplice (plotted as red dots) have square root of profile JSD greater than 0.2 and log coverage greater than 0.5. Setting the genes with square root of profile JSD larger than 0.25 and coverage larger than 5 to have significant difference in profile, DiffSplice achieves a sensitivity of 92%, higher than those of FDM (80%), Cuffdiff with annotation (81%), and Cuffdiff without annotation (58%).

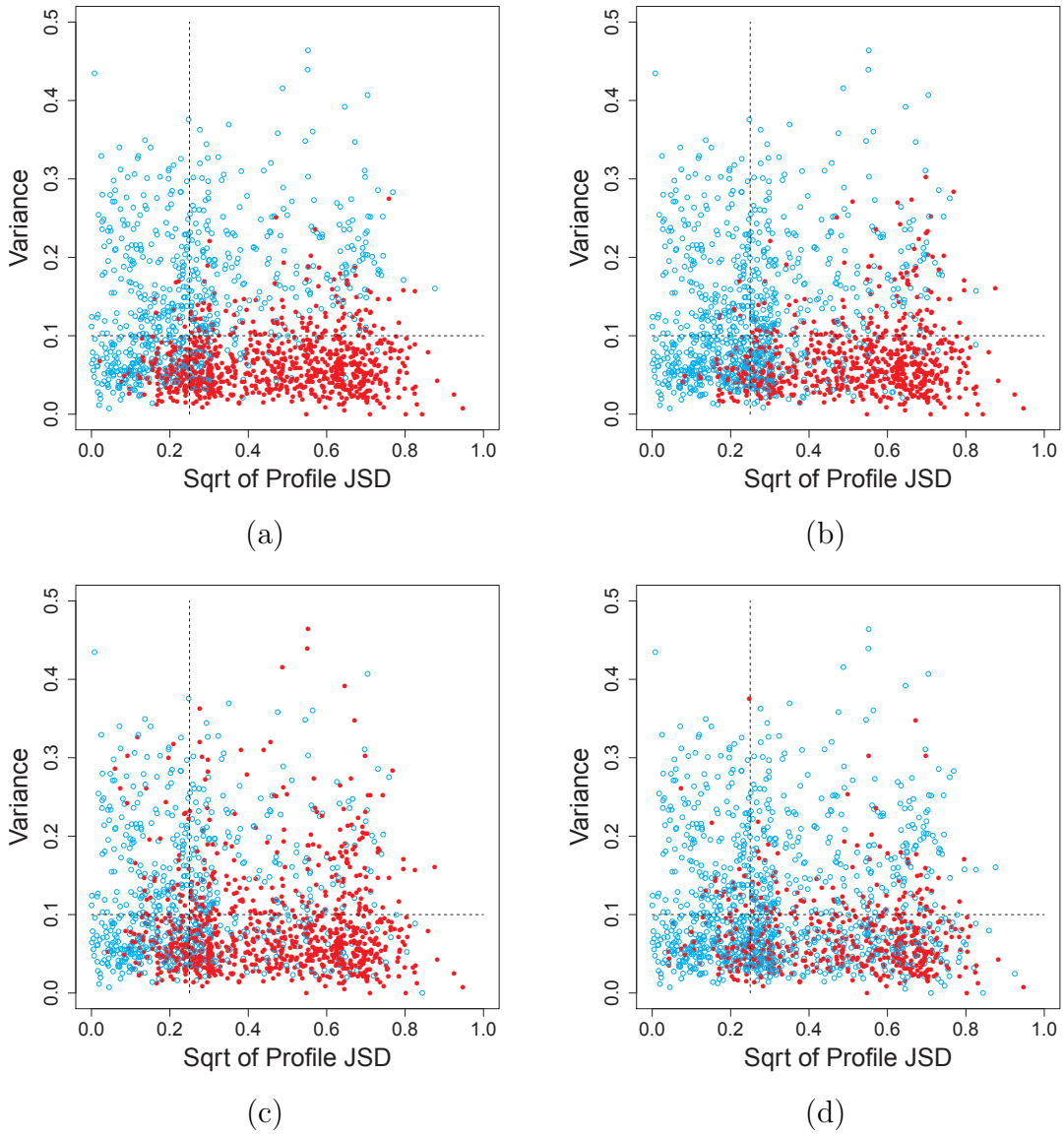


Figure 5.6: Comparison among DiffSplice, FDM, and Cufflinks on simulated dataset of human transcriptome: scatterplot of variance against profile JSD for results of (a) DiffSplice, (b) FDM, (c) Cufflinks with annotation, and (d) Cufflinks without annotation, respectively. Most of the differentially transcribed genes identified by DiffSplice (plotted as red dots) have variance less than 0.1. Setting the genes with square root of profile JSD larger than 0.25 and variance less than 0.1 to have significant difference in profile, DiffSplice reaches a sensitivity of 89%, higher than those of FDM (74%), Cuffdiff with annotation (80%), and Cuffdiff without annotation (40%).

nificantly differentiated genes identified by DiffSplice have large profile JSD (square root of profile JSD > 0.2), showing that DiffSplice correctly captures transcriptional divergences between the two tissues. Along the y-axis, most significant genes identified by DiffSplice have relatively high coverage. This follows the fact that differences present in highly expressed genes are less likely to occur randomly or be introduced by sampling error and hence have higher confidence. We calculate the sensitivity of all four methods at genes that have large profile difference as well as high expression, for example, the region with square root of profile JSD larger than 0.25 and coverage larger than 5. Among the 548 genes in that region (the up-right part), DiffSplice identified 506 genes as significant differences, with a sensitivity of 92% (506 out of 548). This sensitivity is ten percentage points higher than those of FDM (80% or 437 out of 548) and Cuffdiff with annotation (81% or 443 out of 548), and thirty percentage points higher than that of Cuffdiff without annotation (58% or 316 out of 548). To assess the rate of false positives, we further calculate the precision for every method, defined as the proportion of the true significant genes called by each method in all the significant genes called by the method. The precision of DiffSplice (57%) is close to those of FDM (61%) and Cuffdiff without annotation (60%) and is nine percentage points higher than that of Cuffdiff with annotation (48%).

The test statistic of DiffSplice also takes into account the variance of alternative path distributions among the replicates in each group. Figure 5.6 plot the genes coordinated by the square root of its profile JSD and the within-tissue variance of its transcript profile. The genes with significant or insignificant differences are still represented by red dots and blue circles, respectively. Almost all significant genes

identified by DiffSplice have low profile variance compared to profile divergence. We also calculate the sensitivity of all four methods at genes that have large profile difference as well as small within-tissue variance, for example, the region with square root of profile JSD larger than 0.25 and variance less than 0.1. Among the 952 genes in that region (the bottom-right part), DiffSplice identified 849 genes as significant differences, with a sensitivity of 89% (849 out of 952). This sensitivity is fifteen percentage points higher than that of FDM (74% or 705 out of 952), nearly ten percentage points higher than Cuffdiff with annotation (80% or 764 out of 952), and forty percentage points higher than that of Cuffdiff without annotation (49% or 462 out of 952). DiffSplice also has a precision (96%) close to FDM (98%) and clearly higher than Cuffdiff with annotation (82%) and Cuffdiff without annotation (87%).

5.6 Experiments with clinical RNA-seq datasets

5.6.1 Lung differentiation dataset

The human lung airway epithelium lies on the lung-environment interphase, serving as the important physical barrier against invading pathogens. It is composed of various cell types, including ciliated cells, mucus-secretory goblet cells, and basal cells, differentiated from specialized cells in varying numbers. We hypothesized that genes expression changes, including the differential expression of alternative spliced isoforms, are key in the mucociliary cell differentiation and function. Thus, we have sequenced mRNAs from primary human bronchial cells at the early (day 3) and late

(day 35) differentiation stages respectively by high-throughput sequencing. Three biological replicates were used in each group (day 3 versus day 35). Following the manufacturer's instruction, mRNA libraries were made for each sample, and around 28 million 76bp single-end reads were generated from each sample for analysis. The biological findings from this experiment will be presented in another report.

The RNA-seq reads were mapped by MapSplice 1.15.1 [Wang et al., 2010a] to the human reference genome (hg19). About 94% were mapped for each sample. DiffSplice was then performed on these read alignments. Cufflinks+Cuffdiff pipeline (version 1.1.0 with bias correction) was also run on the same read alignments with results both using and not using transcriptome annotation generated for comparison.

As shown in Figure 5.7a, DiffSplice identified 2077 genes that have differential gene expression level between day 3 and day 35 at FDR < 0.01 and requiring the fold change larger than 2 (up-regulated) or less than $\frac{1}{2}$ (down-regulated). This number is similar to the results obtained from the SAM analysis Tusher et al. [2001]. At day 35, 1429 genes were tested to have significantly higher expression level than at day 3, while 648 genes were tested to have significantly lower expression level than at day 3. This observation has indicated active metabolism biogenesis process occurring during the airway epithelium differentiation. At FDR < 0.01, DiffSplice also identified 498 genes exhibiting significant differentiation on alternative transcription. Among them, 109 genes had significantly altered overall gene expression, whereas the rest 389 genes were differentially transcribed while their total gene expression remains at the same level. We randomly selected genes with the inter-group sqrt of JSD > 0.3 for qRT-PCR validation (Supplementary Figure 6). The expression profiles of two validated

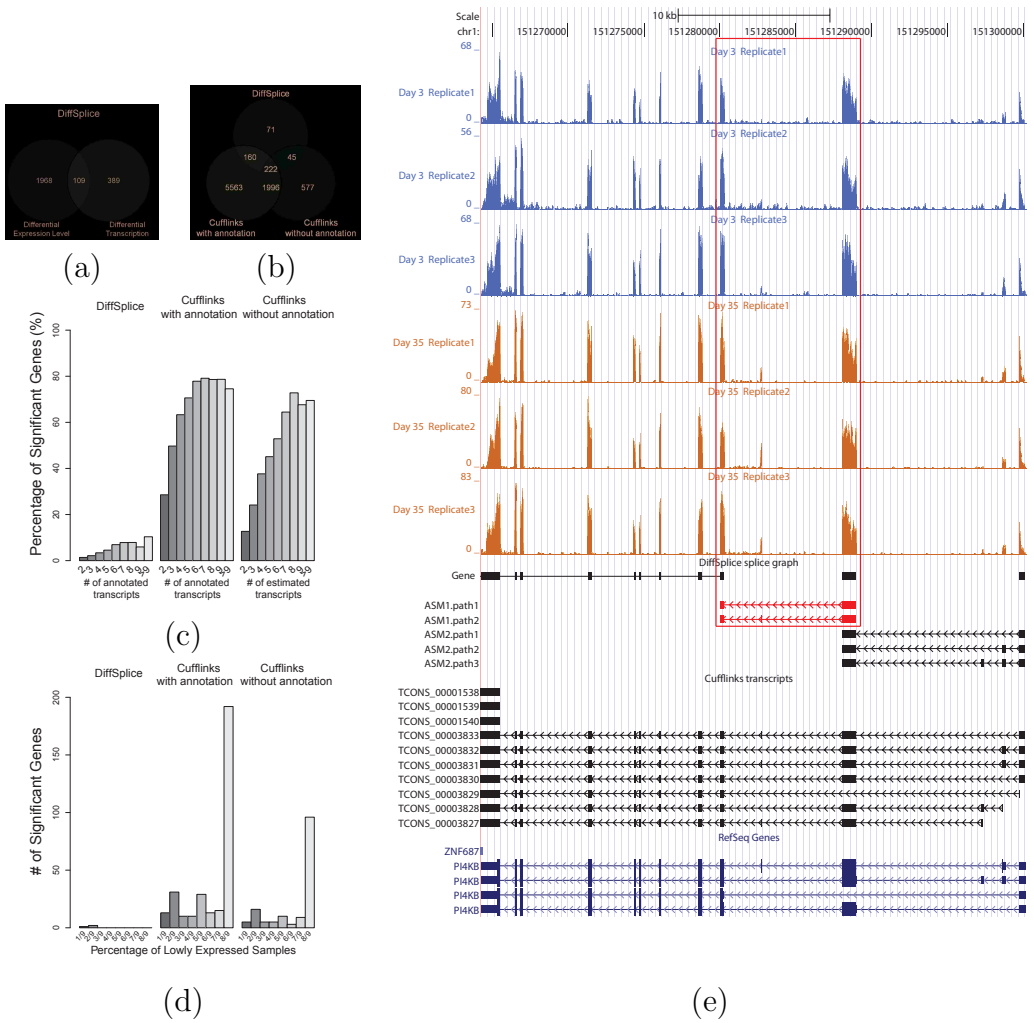


Figure 5.7: Comparison between DiffSplice and Cufflinks on the lung differentiation dataset. **(a)** Differential expression discovered by DiffSplice using MapSplice alignment without annotation. **(b)** Comparison among differentially transcribed genes discovered by DiffSplice, Cufflinks with annotation, and Cufflinks without annotation. **(c)** Percentage of significant genes with differential transcription against number of transcripts. **(d)** Number of significant genes with differential transcription against percentage of samples with gene coverage < 3 in each group. **(e)** Differential transcription in gene PI4KB, identified by DiffSplice but missed by Cufflinks without annotation.

genes TMC5 and LMO7 are included in Figure 5.8 and Supplementary Figure 7.

We compared the differentially transcribed genes identified by DiffSplice and Cufflinks+Cuffdiff. Cufflinks+Cuffdiff with annotation reported over 7000 genes that have significant differential transcription events between day 3 and day 35 while Cufflinks+Cuffdiff without annotation only reported around 3000 genes.In comparison,

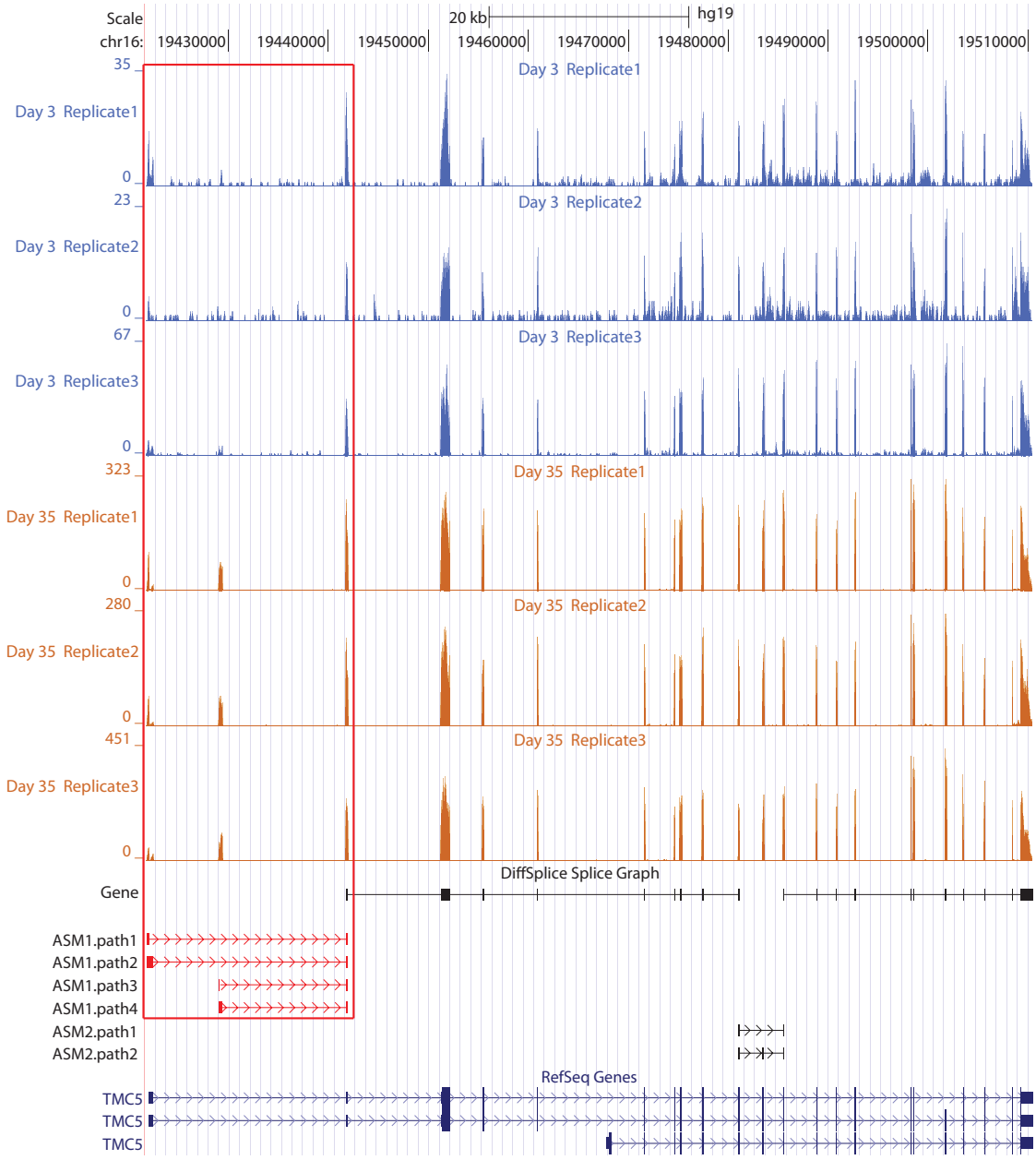
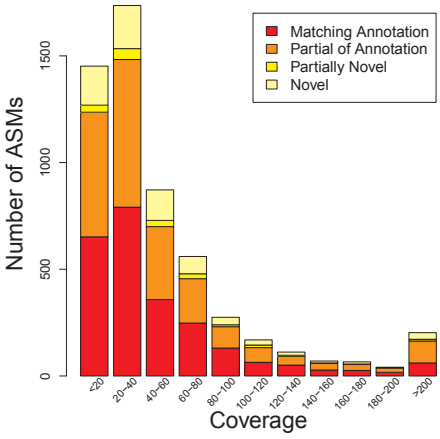
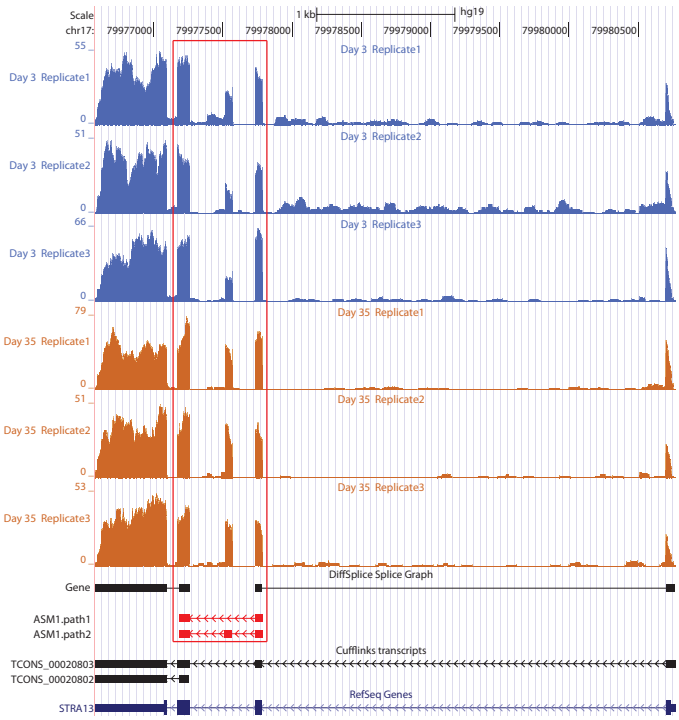


Figure 5.8: Alternative transcription start sites identified by DiffSplice in gene TMC5. The relative expression of isoform passing $ASM1.path4$ increased significantly from day 3 to day 35. The change has been validated by qRT-PCR experiment (Supplementary Figure 6). Meanwhile, the overall gene expression level also significantly increased with a fold change around 11.



(a)



(b)

Figure 5.9: DiffSplice discovers alternative splicing variants present in the data. **(a)** Number of ASMs discovered by DiffSplice at different expression level. Besides around 2,400 ASMs that exactly match annotated ASMs, DiffSplice discovered over 2,000 ASMs where only subsets of annotated splicing variants were present, nearly 200 ASMs with novel splicing variants added to annotated alternative splicing events, and more than 700 ASMs that were completely new to the annotation. **(b)** Novel alternative splicing in gene STRA13, identified by DiffSplice but missed by Cufflinks both with and without annotation. DiffSplice discovered a novel exon in the annotated intron region between the 2nd and the 3rd exon of STRA13. Splice junctions evidenced that the exon was alternatively excluded (path 1) or included (path 2) in transcripts of this gene, and the skipping ratio was tested to have significantly decreased from day 3 to day 35.

DiffSplice reported 498 genes with 77% and 54% overlapped with results of Cufflinks+Cuffdiff with and without annotation respectively. The result is shown as Venn-diagram in Figure 5.7b. Next we detail the major issues from the investigation of the discrepancy.

Effect of transcription complexity. In general, genes with larger number of isoforms tend to have more splicing events, therefore have higher chance to be differentially transcribed. Nevertheless, having the majority of the genes detected to be significantly different indicates a high level of false positive discovery rate. In Figure 5.7c, we divided genes into groups according to the number of isoforms and plotted the percentage of genes detected to be significant as a function of the number of isoforms. As high as 80% of genes with more than 5 isoforms were identified as having significant differential transcription by Cufflinks+Cuffdiff with annotation, and around 50% to 75% of genes with more than 5 known isoforms were identified as significant by Cufflinks+Cuffdiff without annotation. The decreased number of significance called by Cuffdiff without annotation correlates with the typically lesser number of reconstructed transcripts in a gene than the number of annotated transcripts. In contrast, the percentage of genes detected to be differentially transcribed is typically below 10% with DiffSplice with a trend of raising percentage as transcriptome complexity increases.

Transcripts in genes with high transcription complexity are difficult to infer and quantify, requiring a high read coverage to be reliable. Inaccurate transcript inference and/or quantification may not only lead to false positive discovery of the differentially transcribed genes but also miss genes that are truly differentially transcribed. In gene

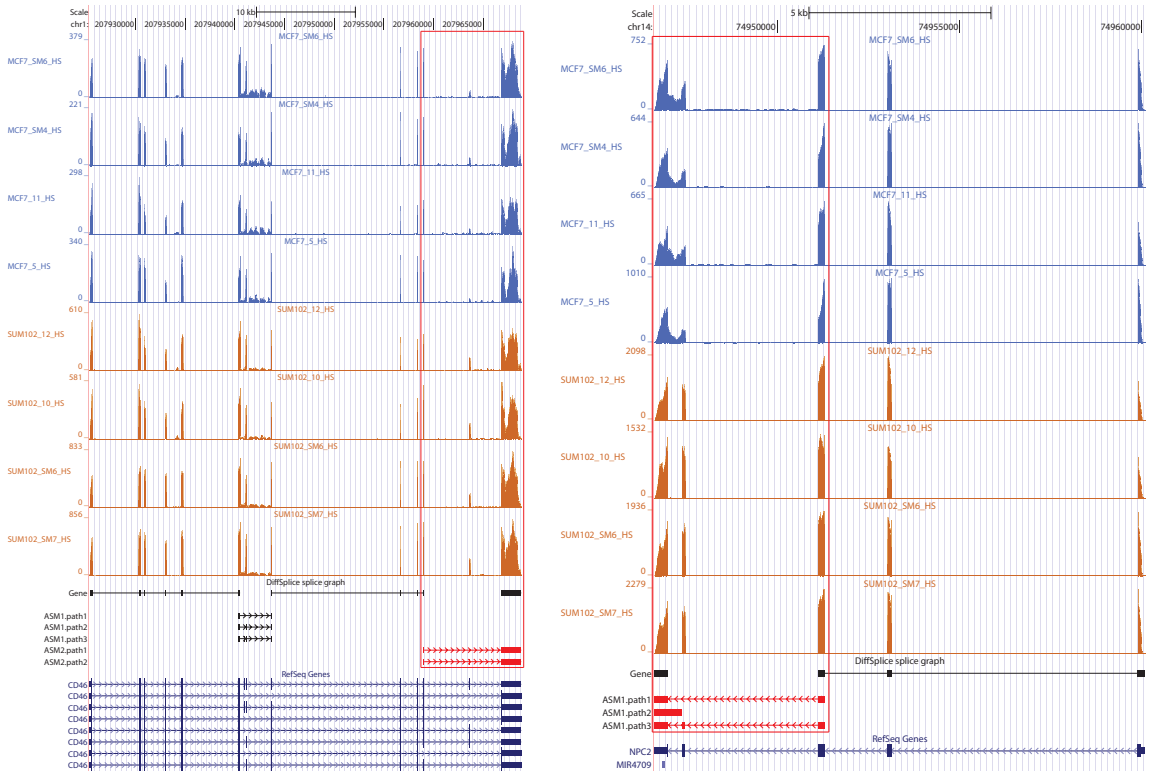
PI4KB, DiffSplice discovered two ASMs as shown in Figure 5.7e. The first ASM starts from the 4th exon (from the 5' end) and ends at the 6th exon, alternatively excluding or including the 5th exon. The second ASM spans from the 1st exons to the 4th exon, alternatively transcribing the 2nd and the 3rd exon. The first ASM was tested to have significant difference in transcription by DiffSplice, which had significantly higher exon skipping ratios at day 3. Without annotation, Cufflinks failed to point out this difference. Cufflinks took the combination of the two alternative splicing events and assembled 7 transcripts, containing 3 spurious transcripts compared to RefSeq annotation. In addition to the inconsistency in assembled transcripts, the estimated transcript abundance by Cufflinks did not reflect the shift on expression. Combining the transcripts that included the 5th exon ($TCONS\_00003827, 00003831, 00003833$), the total expression of the three transcripts was 8.63 at day 3 and 9.27 at day 35 (in RPKM), which did not match the observed increase on the expression of the 5th exon. Also the overall expression of all the 7 assembled transcripts fell from 18.8 (day 3) to 16.6 (day 35), which did not match the observation that the overall expression was actually higher in day 35. In gene TMC5, DiffSplice discovered an alternative transcription start event with 4 alternative start sites and an exon skipping event (Figure 5.8). The alternative start event was tested to have significantly higher abundance of the path $ASM1.path4$ at day 35 (48.9%) than day 3 (14.7%). This finding was consistent with the result of qRT-PCR experiment that the alternative start site corresponding to $ASM1.path4$ had its abundance at day 35 at least twice as high as its abundance at day 3. This gene was also found having differential expression level, with its expression at day 35 more than 10 times higher than that

at day 3.

Effect of coverage and variance in replicates. When determining differential transcription, read coverage needs to be sufficiently high to make reliable inference on the transcript expression. In Figure 5.7d, we plot the number of genes that were called to be significantly different in transcription against the number of samples with exceptionally low expression (*e.g.* gene coverage < 3). The three methods in comparison detect similar percentage of significant genes when the majority of the samples are well expressed. However, Cufflinks+Cuffdiff calls hundreds of genes as significantly differentiated when almost all samples in a group are barely expressed at all.

Besides, Figure 5.7d also indirectly shows high within-group variance among replicates. In testing of differential expressed or transcribed genes, the variance among samples within the same group is expected to be low and should be well controlled. More than 3 out of 9 replicates in one of the comparison groups had extremely low coverage in 269 genes detected by Cufflinks+Cuffdiff with annotation and 128 genes detected by Cufflinks+Cuffdiff without annotation, demonstrating high within-group variance of these genes.

Novel alternative splicing. Since DiffSplice takes only RNA-seq read alignments as input and relies on no annotation, it captures splicing events that are only relevant to the given mRNA samples and has the capability of discovering novel alternative transcripts. We categorize an ASM detected by DiffSplice into 4 types: the ASM exactly matches an annotated ASM; the ASM is a subgraph of an annotated ASM; the ASM partially overlaps with an annotated ASM; the ASM is not found



(a) (b)

Figure 5.10: DiffSplice on the breast cancer dataset. (a) Differential transcription on skipped exon in gene CD46 identified by DiffSplice. DiffSplice discovered two ASMs in this gene. The second ASM that alternatively skipped the 13th exon was tested to have significantly higher skipping ratio in MCF7 samples. This transcriptional difference has been validated by qRT-PCR experiment. (b) Differential transcription on retained intron in gene NPC2 identified by DiffSplice. The exon-skipping event spanning the left three exons was tested to have significantly higher skipping ratio in MCF7 samples. The nested intron-retention in the left two exons was also tested to have significantly higher ratio of retaining the intron in MCF7 samples. The differential transcription in the intron-retention event has been validated by qRT-PCR experiment.

in the annotation. The histogram of each category at varying coverage is shown in Figure 5.9a. The ASMs detected by DiffSplice show high consistency with those generated from known annotation. Among the totally 5,556 ASMs found be DiffSplice, 2,426 ASMs matched an annotated ASM, 2,219 ASMs were subsets of annotated ASMs. Besides the alternative splicing events present in annotation, we found 174 ASMs with novel paths added to annotated ASMs and 736 novel ASMs. For example, we discovered a novel exon in gene STRA13, located between the second and the third exon in the RefSeq annotation (Figure 5.9b). This exon was discovered as differentially skipped between day 3 (50% skipping ratio) and day 35 (30% skipping ratio). Because the exon-skipping event in STRA13 is not present in the transcriptome annotation, Cufflinks with annotation did not capture the difference. Cufflinks without annotation falsely initiated a transcript from the third annotated exon and did not detect the event either.

5.6.2 Breast cancer MCF7-SUM102 dataset

We further applied DiffSplice to the RNA-seq datasets generated from two breast cancer cell lines, MCF7 and SUM102 [Singh et al., 2011]. Each cell line group comprises of 4 technical replicates and about 80 million 100bp single-ended reads were sequenced for each replicate. Flow Difference Metric (FDM) was originally applied to these datasets to detect genes that might have differentially transcribed without usage of transcriptome annotation information [Singh et al., 2011]. At FDR < 0.01, DiffSplice identified 6103 genes with significant difference on expression level and 2507 genes with significant difference on transcription between the two cell lines, includ-

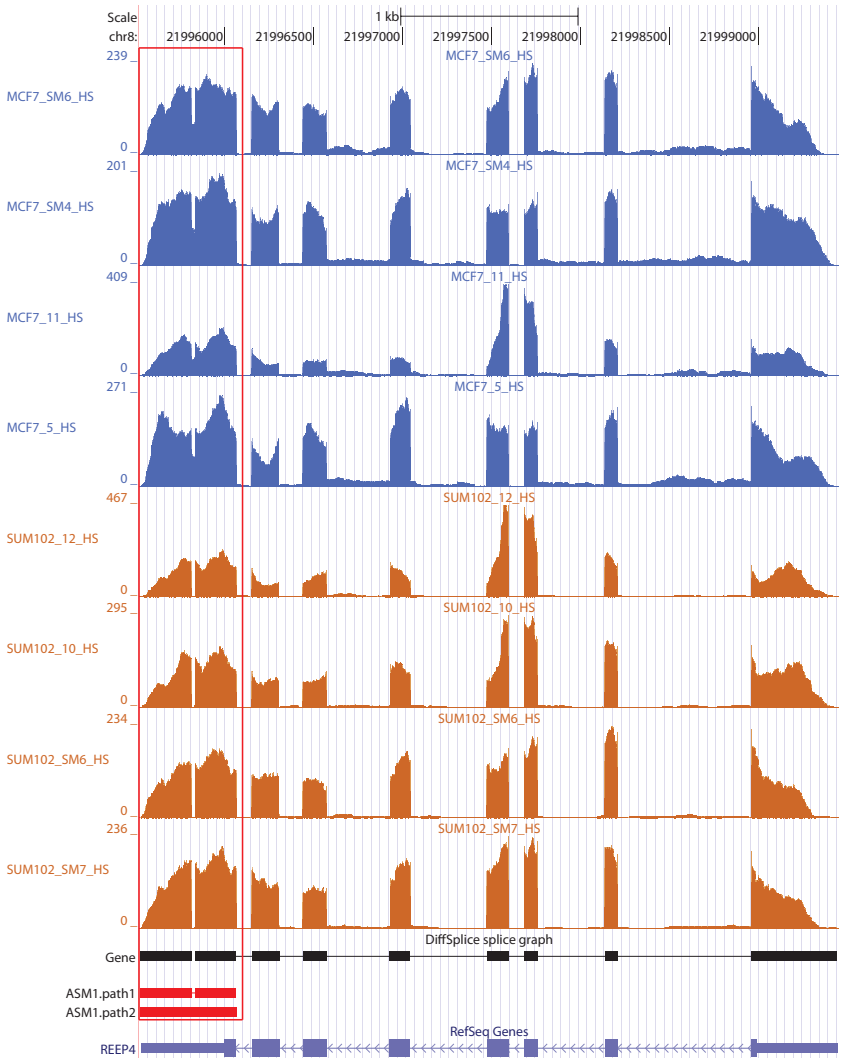


Figure 5.11: A novel deletion in gene REEP4 found differentially transcribed between SUM102 and MCF7 by DiffSplice. In SUM102, 19 bases were deleted in almost all transcripts compared to the reference genome. In MCF7, the deletion was only present in approximately half of the transcripts. This novel deletion has been validated through resequencing.

ing 1353 genes with both differences. For genes that were differentially transcribed, DiffSplice had 955 (38.1%) shared with those discovered by FDM (Supplementary Section 6.3).

DiffSplice successfully identified the two genes CD46 (Figure 5.10a) and NPC2 (Figure 5.10b) that were originally validated by qRT-PCR in FDM paper. However, unlike FDM, DiffSplice directly pinpoints the location of alternative splicing events that are differentially expressed, consistently with those chosen for the qRT-PCR validation. For example, in the exon-skipping event found in CD46 (Figure 5.10a), the averaged estimated proportion of the path that included the 13th exon (chr1:207963598-207963690) was 34.7% in the SUM102 group and 13.9% in the MCF7 group. This result was consistent with the observation in the qRT-PCR experiment that the skipped exon had more than two fold higher expression level in SUM102 than in MCF7. In gene NPC2, DiffSplice discovered two alternative splicing events, one nested within the other (Figure 5.10b). The intron retention occurs between the last two exons was found present primarily only in MCF7 samples. This ASM was further nested in a larger exon-skipping event spanning the last 3 exons, where the 2nd exon was alternatively spliced with a significantly lower skipping ratio in SUM102 samples. The first intron-retention event was picked for qRT-PCR validation. The averaged estimated proportion of the path that retained the intron (chr14:74946992-74947388) was 0.5% in the SUM102 group and 17.9% in the MCF7 group, consistent with the experimental observation that the retained intron had at least ten fold higher expression level in MCF7 than in SUM102.

Besides alternatives spliced events, DiffSplice can be generalized to detect struc-

tural variations whose presence is different across two comparison groups. 42 genes were detected to have a small insertion/deletion that varies between MCF7 and SUM102. As shown in Fig. 5.11, a 19-base novel deletion was discovered in the last exon of gene REEP4. The averaged estimated proportion of the path that included the deletion was over 99.2% in SUM102 samples. The estimated proportion of the deletion fell to 49.9% as turning to MCF7 group. We directly resequenced the genomic DNA as well as the cDNA derived from the mRNA of the cell lines and validated this novel deletion. These deletions evidenced the genomic variation present in cancer cell lines and may contribute to prognostic differences together with other differential expression events.

Copyright © Yin Hu, 2013.

Chapter 6 Differential Splicing Analysis on Large-scale Datasets

6.1 Introduction

The RNA-seq technologies may comprehensively and accurately profile the transcriptome of a cell at a specific condition. However, detecting the transcriptomic characteristics of one individual or a small group of individuals may not lead to improved understanding of cell functions such as cell differentiation and progression and genetic diseases such as cancers. The sample size is the key to the expansion of knowledge and its population-wide application.

Fortunately, the price of sequencing has dropped. Nowadays, an RNA-seq experiment typically costs less than $1,000, comparable to the cost of microarray. This allows wider applications of sequencing technologies. Several large-scale projects have been initiated in order to decipher various types of human diseases as well as the function of human genome, such as TCGA (The Cancer Genome Atlas), ICGC (International Cancer Genome Consortium), CCLE (The Cancer Cell Line Encyclopedia) and PCBC (Progenitor Cell Biology Consortium). The TCGA project, for example, was launched by the National Cancer Institute and the National Human Genome Research Institute, both part of the National Institute of Health, in 2006. This project aims to comprehensively sequence, characterize and analyze more than 20 types of cancer, types with poor prognosis and overall public health impact. A total of more than 4,000 samples have been sequenced, providing an unprecedented opportunity to

discover the mechanism of cancer.

The massive amount of data is also changing the biomedical research from hypothesis-driven to data-driven. Analyses are carried out directly on the raw data, with partial or little reliance on existing reference. The exploration of biological insights is being greatly accelerated.

Despite these exciting advancement of sequencing technologies, the scale and complexity of the data have imposed three central difficulties in transforming the massive raw data to potential biological findings.

Most directly, the increased sampling depth and sample size lead to trillions of reads, which require significant computing time and storage. For example, the RNA-seq data of TCGA consists of $\sim$ 5000 samples, more than 1 trillion reads and approximately 50 TeraBytes of binary files. A typical pipeline, such as Cufflinks, may take one day to analyze one sample, making it expensive to process the data for even once.

More importantly, adding more samples not only requires more computing resources, but also leads to (sometimes exponentially) increased complexity. For example, a large RNA-seq dataset may introduce a list of putative splice junctions several folds longer than the list of annotated junctions, further resulting in tens or even hundreds of possible transcripts in a gene. The later transcript quantification and differential test steps then become questionable.

Lastly, the statistical power brought by the large sample size needs further exploration. Current methods for differential transcription typically conduct comparisons between two samples, *e.g.*, one normal versus one tumor or one control versus one

drugged. Some methods perform analysis between two groups of samples, but the sample size in each group is typically less than 20. The heterogeneity in the samples may also need analysis beyond the classical group-wise hypothesis tests.

We have developed a novel pipeline for the differential transcription analysis on large-scale RNA-seq datasets. This pipeline utilizes a joint analysis model that summarizes all samples with a single graph, which can leverage information from all samples. A suite of components have been designed to target the challenges above, including a preprocessing step that accelerates the construction of the graph and alleviates the data storage burden, approaches of splice junction filtering and exon boundary detection that clean the noise and insignificant signals in the data, a multi-group test statistic and a clustering approach that detects transcription patterns across sample groups. We have demonstrated the differential splicing analysis on a set of 819 RNA-seq samples, which came from the breast cancer (BRCA) analysis of the TCGA project. These samples constitute a large dataset with a total of 6 TeraBytes of data in binary format, 5-10 GigaBytes per sample. Every sample contains 120M to 250M 2×50bp paired-end reads. These samples have also been clinically classified as normal (91) and tumor (728), with 5 tumor subtypes determined by clinical characteristics and gene expression: Basal (123), HER2 (60), LumA (359), LumB (170) and normal-like (16).

6.2 The joint transcriptome analysis of all samples

Two possible pipelines may be built upon current single-sample transcriptome analysis methods. Relying on transcript annotation, a transcript abundance quantification

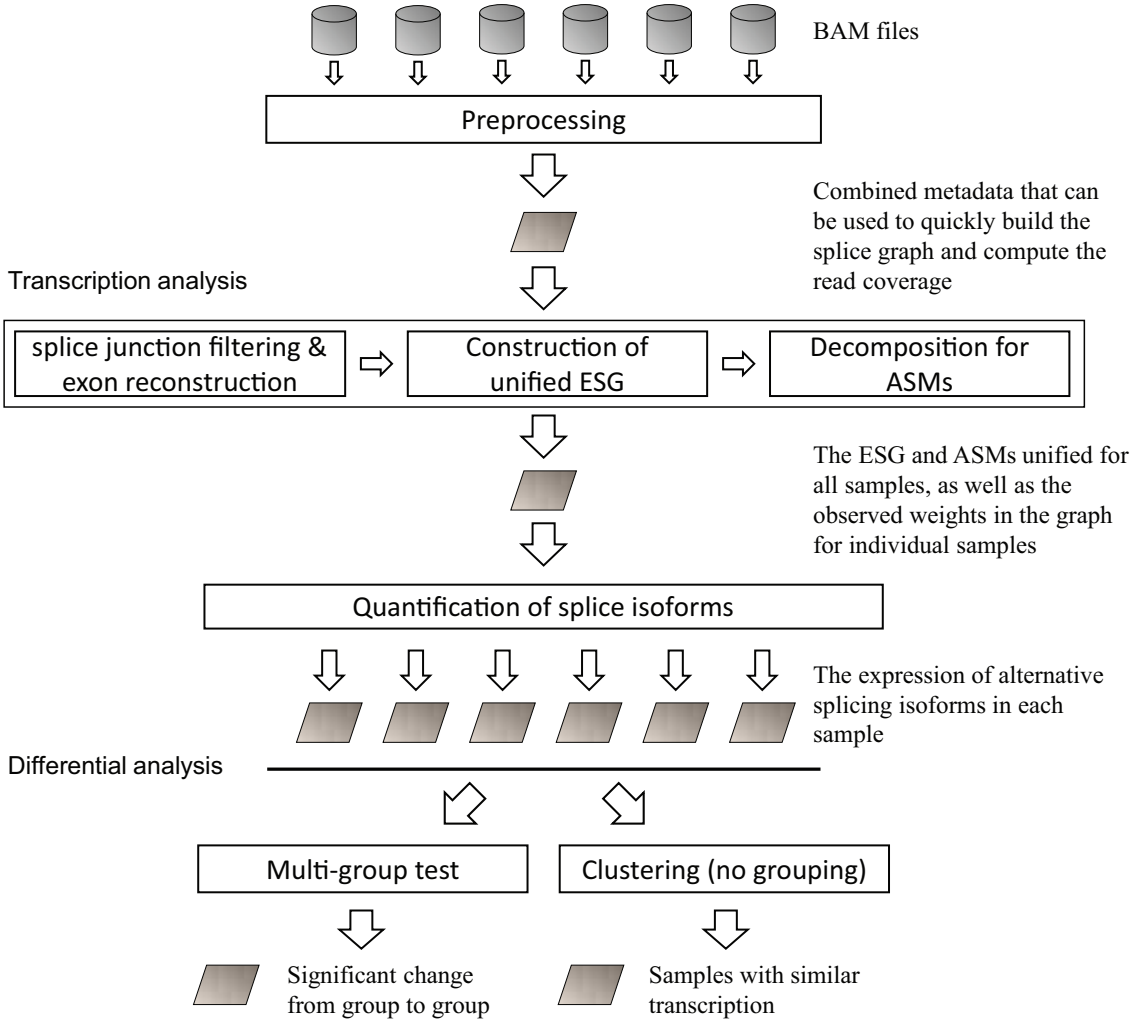


Figure 6.1: An overview of the scalable pipeline for alternative splicing and differential analysis on large-scale RNA-seq datasets. The expression information contained in the read alignments (BAM files) is combined via a distributed preprocessing step. The transcription analysis is then conducted on the unified ESG that represents all samples. The expression of splice isoforms is quantified for each sample. Differential analysis is lastly applied on the basis of the estimated isoform expression.

method (such as RSEM [Li et al., 2010a]) can be applied on every sample, followed by the application of a differential test method (such as DESeq [Anders and Huber, 2010]) to determine the transcripts whose expression has changed significantly. This approach, however, prevents the discovery of novel transcripts and may cause inaccurate quantification due to expression of unannotated isoforms. Therefore, it is not ideal for the discovery of disease biomarkers. Alternatively, transcripts present in each sample, including novel transcripts, can be reconstructed *ab initio* using methods such as Cufflinks [Trapnell et al., 2010]. Nonetheless, transcripts that have low expression in some sample may be missed in the sample due to computational regularization. None of these two approaches consolidates the transcriptome-wide discovery using information benefited from a large RNA-seq sample size.

Computationally, reconstruction and quantification of full transcripts are unsolved problems, difficult and inaccurate because of the inadequate read length. The quantification may be underdetermined, meaning that a unique solution to the transcript abundance does not exist.

To overcome the above shortages, we have developed an approach for integral differential splicing analysis that considers all samples jointly. The core algorithm summarizes the expression information, such as splice junctions and read coverage, of every individual sample together as a unified splice graph (Figure 6.1). This gives more accurate detection of exon boundaries and splice junctions, and allows the examination of all alternative splicing patterns in one graph. Efficient algorithms for alternative splicing detection can then be applied on the unified splice graph, without the need to repeat on every single sample. By focusing on the alternative splicing

events, the method circumvents the need of transcript-level reconstruction and estimation, thus may provide more accurate inference than directly using transcripts as basis. Lastly, the expression in each sample is estimated individually, and different samples can be compared on the same graph with their sample-specific weights. This procedure will highlight differential splicing events, the alternative splicing events that exhibit significantly different splicing ratios between different samples or between different subtypes.

6.2.1 Preprocessing of the alignment files

The read coverage on the genome is necessary toward the construction of a splice graph. However, parsing the read alignments one by one is time-consuming, because every read accounts for a separate calculation, further complicated by spliced alignments. Importantly, these expensive calculations need to be repeated every time the data is re-analyzed, making any transcription-level analysis a considerable burden.

In order to address this, we introduce a preprocessing step to summarize the raw data with less records, in a format that simplifies the computation of read coverage. This step is done by splitting an read alignment into exonic segments that represent continuous sequences covered by the reads and splice junctions that connect the exonic segments. Same entries, such as same splice junctions and exonic segments that cover a same short exon, can be merged into one entry, associated with the count of its presence.

We have designed a MapReduce scheme to parallelize the preprocessing step:

Mapper Every worker node parses the read alignments of one sample at a time, extracts the exonic segments and splice junctions of the sample, and maintains a count table for every unique entry;

Reducer The master node combines the count tables of all samples.

Under this scheme, samples can be parsed simultaneously at distributed computing cores. The information from all samples, summarized as a count table for each sample, will be combined using a linear-time merge sort-like algorithm.

Specifically for the BRCA dataset, it costs > 20 days for directly parsing all read alignment files, with 20 computing cores. Using the preprocessing step, the same task can be completed in < 7 days under the same computing power. Notably, it is not necessary to repeat the preprocessing step, which enables fast analyses for < 0.5 day. At the same time, the storage can be reduced from $\sim$ 6TB to $\sim$ 730GB, approximately 10% of its original size.

6.2.2 The filtering of spurious splice junctions

The alignments of the RNA-seq short reads often suggest a large number of splice junctions. For example, a total of 1,350,669 putative splice junctions have been detected by the reads in the BRCA dataset, including 222,248 (16.5%) in existing annotation and 1,128,121 (83.5%) not in annotation. As a comparison, there are only The unannotated junctions may imply novel splice isoforms and novel mRNA transcripts that have not been cataloged. However, they may also result from sequencing errors or mapping errors.

It is not trivial to separate true junctions from those falsely detected. The most informative and commonly used feature for the measurement of a splice junction's existence is the number of read alignments that span the junction, denoted as the number of spanning reads or simply the support. For example, splice junctions that have high support (*e.g.*, > 10) tend to be known. Splice junctions that have low support, on the other hand, mostly have not been cataloged in existing knowledge. Nonetheless, a specific cutoff on the support is not sufficient, because the expression of different genes can vary in magnitudes and it is often not practical to locate a "good" cutoff that balances the specificity for highly-expressed genes and the sensitivity for lowly-expressed genes. More importantly, even when a satisfying filter can be achieved, the threshold only applies to the specific dataset.

We have developed an adaptive filter for putative splice junctions, which utilizes four metrics to evaluate the confidence and significance of a splice junction detected by the RNA-seq read alignments. The metrics are defined based on both the sequences associated with a splice junction and its relative expression as compared to the gene.

Anchor complexity The genome is a 3 billion bases-long sequence of just 4 different nucleotides. Therefore, different regions of the genome may have highly similar sequence, especially for cases such as pseudo-genes. While the ambiguity of alignment may be relatively low for an entire read, which is typically longer than 50 bases, there may exist multiple places with similar sequences for the alignment of a read fraction due to splice (an anchor of the splice). Therefore, a splice junction may be spurious if all the spliced alignments that span the junction have short anchors, *e.g.*, less than 15 base pairs. In such case, the sequence of the anchor can then be

compared to the sequences of the genome, using a method such as BLAT [Kent, 2002]. Splice junctions with only low complexity anchors detected may be filtered.

Expression of anchor exons A splice junction conjoins two separated exons when the gene is transcribed. Therefore, both anchor exons of a splice junction should have sufficient expression in order to evidence the expression of the corresponding transcript. In our application, an anchor should have a minimum read coverage of 1 in order to be distinguished from noise.

Expressed sample percentage The support of a junction is the most direct evidence of the existence of the junction in the sample. However, the absolute support is often not comparable across genes and may not be used as a filter cutoff, because genes may differ greatly in their expression level. In genes with low expression, true and important splice junctions may have few reads spanned. Therefore, we calculate the *relative support*, to indicate the relative expression of a splice junction as compared to the expression of the exons connected by it:

$$RS(j) = \frac{\text{support of } j}{\min(\text{read coverage of } e_5^j,\ \text{read coverage of } e_3^j)}, \tag{6.1}$$

where e_5^j and e_3^j are the donor and acceptor exons of j, respectively. Then the expression of a splice junction in a sample can be defined with a cutoff on the relative expression τ. The expressed sample percentage given τ, $ESP - \tau$, is calculated as

$$ESP - \tau = \frac{\#\text{ of samples in which support of } j > \tau}{\text{Total } \#\text{ of samples}}, \tag{6.2}$$

which is equivalent to $1 - percentile(\tau)$.

Top α% sample expression. At last, some splice junctions may not be present in the majority of samples, but they may have high expression in a small portion of

samples. These junctions may encode important transcript isoforms unique to a small group of samples. Therefore, we further calculate the $(1-\alpha)$-percentile of the relative expression in all samples, denoted as $TOP-\alpha$, to gauge the relative expression of the splice junction in the top $\alpha\%$ highly expressed samples.

Thereafter, we construct a splice junction filter that also leverages the information of gene splice structure.

1. *Form a trusted list.* The anchor complexity and anchor exon expression are used to distinguish splice junctions that have high probability to be real. The metrics $ESP-\tau$ and $TOP-\alpha$ are then to rank the splice junctions considered true and select a trusted list with significant relative expression.

2. *Construct a reference gene model.* The selected splice junctions are used to build a basic splice graph to represent the gene model. Exons and introns are recognized according to the position of the splice junctions, and the read coverage is calculated.

3. *Aggregate toward a final list* The splice junctions not selected in the first step tend to have lower confidence or significance. The gene model obtained in step 2 is used to assist consolidating the final splice list with the unselected splices. In particular, splice junctions that connect existing splice sites in the reference gene model have higher probability to be real. Splice junctions may require extra evidence (higher anchor complexity/expression and relative expression) if they do not readily fit the reference gene model, *e.g.*, junctions connected to an intronic sequence and long-range junctions.

After filtering, 184,663 (13.7%) splice junctions have been kept for the construction of the transcriptome. Within the junctions kept, 161,896 (87.8%) are in junction annotation, demonstrating the effectiveness of the filtering process.

6.2.3 Identification of exon boundary

However, it is difficult to specify the expression cutoff without knowledge of the read coverage of the mixed exons.

Through wavelet transformation, we have been able to distinguish the exonic signal from the noise signal automatically. Every mixed region is first identified by recognizing an exon connected by splice junctions of different strands, which indicates that two oppositely stranded genes are falsely joined by noise. The read coverage of the mixed region is extracted. A one-layer wavelet decomposition is performed on the read coverage signal, leading to a high frequency component (the detail component) that magnifies the difference between coverage on exonic nucleotides and that from noise. Such difference can be captured by calculating the variance ratio at a cutting base pt,

$$VR(pt) = \frac{Var(coverage[0, pt])}{Var(coverage[pt + 1, n])}. \tag{6.3}$$

The exon boundaries of the mixed region are then determined by finding the two bases with the highest variance ratio.

The time complexity of this procedure is linear to the number of bases in the region, *i.e.*, the length of the signal. In the BRCA dataset, we have successfully processed 1,707 mixed regions.

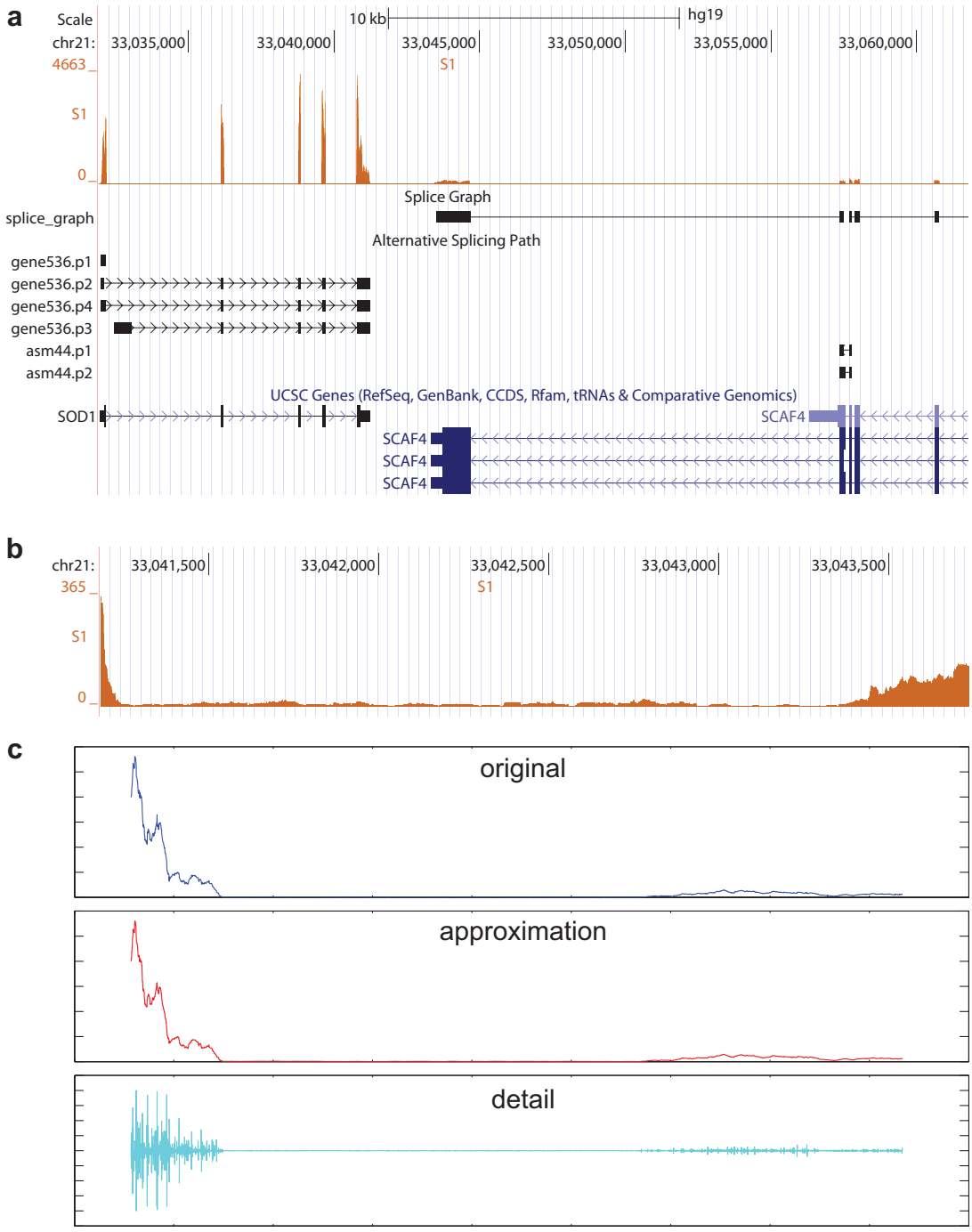


Figure 6.2: The recognition and removal of read coverage noise through wavelet decomposition. **(a)** The read coverage of gene SOD1 and SCAF4 of a BRCA sample. The tracks labeled splice graph and alternative splicing paths show the splice graph reconstructed by DiffSplice. **(b)** The read coverage in the region that connects the two adjacent exons of SOD1 and SCAF4. The read coverage outside the exons is high (> 20), making it difficult to recognize the boundaries of the exons. **(c)** The wavelet decomposition of the read coverage signal. The variance of the detail component distinguishes the exonic regions from the background.

6.2.4 Weight of the splice graph

The splice graph depicts the possible ways the exonic sequences of a gene can get transcribed. Every graph path representing a transcript, the start/end vertex indicates the transcription start/termination site, and the graph edges through which vertices in the path are traversed identify the exon composition of the transcript.

Further, the expression of the transcript is reflected by the number of reads sampled on each exon. The expression information is represented by the weights in the splice graph. In particular, the vertex weight represents the averaged read coverage on an exon, whereas the edge weight represents the number of spanning reads of a splice junction.

Because the number of reads sampled in each sample may be different, the expression in each sample needs to be normalized. We normalize the samples by the total number of reads in each sample, which is an estimate for the library size. For datasets with less heterogeneity, normalization techniques such as upper-quartile normalization [Bullard et al., 2010] and median normalization may also be applied, under the assumption that the genes in the specific quartiles have the same median expression level in different samples.

6.2.5 Transcription analysis on the unified splice graph

The preprocessing together with the splice junction filter and the exon boundary detector allow fast and accurate reconstruction of a unified splice graph that represents all samples. The transcription analysis is then performed on this graph using

algorithms discussed in previous chapters. Specifically, the genome-wide ESG is first decomposed into a hierarchy of ASMs, which capture all alternative splicing events in the data. Every sample then shares the same transcription model summarized by the graph, but with different weights determined by the sample-specific expression. The expression of each alternative splicing isoform is then estimated and will be used as the basis of downstream differential analysis.

6.3 Cluster analysis of transcription in alternative splicing

Traditionally, differential transcription analyses are defined as the comparison of transcriptome between two conditions, *e.g.*, a group of samples from diseased individuals versus a control group of samples from normal individuals, or that among multiple classes, *e.g.*, samples from different tissues of a human body or samples from different subtype groups of breast cancer patients. The basic assumption of these traditional group-wise analyses is the homogeneity of the samples pre-specified in each condition group, that is, all samples in a class should follow some "population distribution". These population distributions are often described by single-mode probability density functions with population-wise parameters. For example, in the context of gene expression, the expression level of a gene in samples of a condition may be modeled as a log-normal distribution with a population mean and a population variance. For any sample in this condition, the scaled expression level of this gene after taking the logarithm is expected to center closely around the logarithm mean. However in large-scale clinical data, samples pre-grouped may have prominent variations which will make the assumption of a population distribution questioned. To calculate the

difference between two sample groups and to evaluate the statistical significance may become problematic because of the biases on the mean and variance. Furthermore, if the samples in a group exhibit more than one pattern, the consistency in transcription of these samples may be neutralized and overlooked if all samples in the group are forced to be treated as the same.

Therefore, an approach not depending on presumed grouping will be highly valuable for complex large-scale clinical data. As a unsupervised approach, cluster analyses may serve as a reasonable solution for these datasets.

6.3.1 Related work

Application of clustering techniques is not new in the detection of expression patterns of genes across samples or datasets. Nonetheless previous work has mostly focused on inference in the co-regulation of expression at the gene level.

Co-expression network inference: Genes are molecular units in living organisms which carry the important biological information encoding proteins. The interaction among the genes are complicated, some of which may regulate the behavior of the others. To understand the regulatory interaction mechanisms, a gene network is usually adopted to represent the relationship among the genes. Each node corresponds to a gene of interest and each node is associated with an expression vector denoting its expression under an order list of conditions. The goal is to infer the regulation through the clustering of the gene expression data over a range of conditions on this gene network. The "distance" between a pair of genes is measured between the expression vectors of these genes using certain metric. Afterwards, clus-

tering methods may be applied on the weighted gene networks to find tight clusters, each representing a set of genes co-expressed under the list of conditions. A plenty of work has been published using co-expression network to inference the co-regulation on gene level. For example, Zhang et al. [2012] implemented weighted network mining algorithm called quasi-clique merger and used microarray gene expression datasets from multiple types of cancer to identify the gene co-expression clusters with enriched functions.

bi-clustering: Another strategy to cluster the gene expression values in multiple samples is bi-clustering, a data mining technique which allows simultaneous clustering of the rows and columns of a matrix. Given a set of genes, each with a vector of values representing its expression under a list of conditions (samples), a matrix can be constructed to represent the expression data. Each gene corresponds to one row and each condition to one column. The ij-th entry in the matrix represents the expression level of the i-th gene under the j-th condition. Finding co-expression genes using bi-clustering methods is equivalent to discovering homogenous patterns (blocks) on the matrix. Specific algorithms have been developed to seek for bi-clusters with particular properties, such as bi-clusters with constant values, bi-clusters with constant values on rows (genes) or columns (conditions), bi-clusters with coherent values and bi-clusters with coherent evolutions, to name the major categories [Madeira and Oliveira, 2004].

co-splicing clustering: With the rapid development of RNA-seq technology, one can go beyond the gene-level analysis to alternative splicing events investigation. Alternative splicing allows for a single gene to code for multiple proteins which greatly enriches the protein diversities. For example, in human, about 20,000 protein-coding

genes encode over 75,000 isoforms resulting from alternative splicing events. Also abnormal variations in splicing may be associated with diseases like cancer. Therefore, instead of gene-level co-regulation analysis, it is more interesting to look at the splicing variants and see their behaviors across conditions. One work published recently turned their focus into identifying co-splicing clusters on a set of cassette exons across multiple samples [Dai et al., 2012]. They assume that the co-expressed cassette exons are regulated by the same splicing factors. In this paper, the authors first collected all the cassette exons from the annotation database, then built a 3-layer network to represent the relationship among all the features. The first two layers summarize the relationship between any two features under one condition and the third layer represents different conditions. By finding homogeneous patters across the 3rd layer, this method could get a set of co-expressed cassette exons.

The existing techniques are limited in the scope of classical mining on distance matrices, for the identification of sub-matrices sharing similar values or baring predefined patterns. Every gene in every sample is represented by a feature of a single real value, like the expression level or the skipping ratio. However, in transcription-level clustering it is difficult to find a representative real value for the profiling of transcription, because different genes have different sets and numbers of transcripts. Hence it is hard to define a consistent measurement for the direction and magnitude of change in transcription that may be compared across genes and samples. Therefore, the following content in this chapter aims to propose a new generalized cluster scheme for transcription-level clustering.

6.3.2 Features for clustering

The cluster analysis is performed after the DiffSplice pipeline, on one ASMs at a time. Derived from the unified splice graph that contains splicing variants present in all samples, an ASM has the same set of alternative transcription paths in every sample, but with different expression profiles. In the group-wise statistical tests for differential transcription, the proportions of the alternative paths in the ASM are extracted and the divergence is calculated among the distributions of abundance on the paths in the two groups' samples. Taking the proportions will free the information of the absolute expression levels, which will be examined separately by the tests for different gene expression. Consistent changes in the percentage composition of the splicing isoform will be selected by the tests, even the expression level of the ASM varies in the samples. In the clustering analysis, samples shall be grouped for an ASM if they exhibit consistent transcription profile in the alternative splicing, both in splicing isoforms' composition and their expression level. Therefore, we will use the absolute expression level of the alternative transcription paths in an ASM as the features for clustering.

Consider an ASM A with l alternative transcription paths $p_1, p_2, \cdots, p_l$. Let $C_s(p_i)$ denote the expression level of path p_i in sample s, measured by the read coverage. The expression feature is then defined as the logarithm-normalized value, $\log C_s(p_i)$. The feature vector of a sample s in ASM A is then the l-dimensional real vector $f_A(s) = [C_s(p_1), C_s(p_2), \cdots, C_s(p_l)]^T$. In practice, a small value is added to the expression level in order to prevent taking logarithm on a 0 value, such as 10^{-3}.

6.3.3 Hierarchical clustering with Mahalanobis Distance

For agglomerative clustering, at each level, we need to decide which pairs of clusters should be merged. Therefore, a similarity metric is required for computing distances between any two clusters. We adopt the *Mahalanobis distance* in order to take the variance of the data points within the cluster into consideration.

Given two random vectors $\vec{x} = [x_1, x_2, \cdots, x_n]^T$ and $\vec{y} = [x_1, x_2, \cdots, x_n]^T$, the mean of $\vec{x}$ is $\mu = \sum_{i=1}^{n} x_i$ and covariance matrix of $\vec{x}$ is: $S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$. The Mahalanobis distance between $\vec{x}$ and $\vec{y}$ is defined as:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{y} - \mu)^T S^{-1} (\vec{y} - \mu)}. \tag{6.4}$$

We then use the agglomerative clustering, a bottom-up scheme, to gradually aggregate all samples into clusters, with the average distance criterion. Because the Mahalanobis distance can be viewed as the distance between cluster means normalized by the covariance, we separate two clusters if their distance is > 3, instead of specifying the maximum number of cluster prior to the analysis. The cutoff 3 is motivated by the 3σ rule for the Gaussian distribution, *i.e.*, about 99.7% of values drawn from a Gaussian distribution are within 3 standard deviations from the mean.

6.3.4 Information score

The derived clusters are of different sizes and have various compositions of samples from each subtype group. In Figure 6.3 we show two clusters that may encode different degrees of information. The middle panel of Figure 6.3 shows a cluster C_1 consisting of 568 samples. The percentage of samples coming from each subtype

group is $\rho(C_1) = [\frac{93}{568}, \frac{46}{568}, \frac{280}{568}, \frac{136}{568}, \frac{13}{568}]^T = [0.16, 0.08, 0.49, 0.24, 0.02]^T$, very similar to the subtype distribution of the entire data $\rho(data) = \frac{123}{728}, \frac{60}{728}, \frac{359}{728}, \frac{170}{728}, \frac{16}{728}]^T = [0.17, 0.08, 0.49, 0.23, 0.02]^T$ which is shown on the left panel. Then the evidence is weak for arguing the non-randomness of composing the cluster, indicating that the differential transcription exhibited by the clustered samples on the corresponding ASM may be not informative for classifying the subtypes. On the other hand, the cluster C_2 shown on the right panel has a subtype distribution highly divergent to that of the entire data, $\rho(C_2) = [\frac{46}{57}, \frac{9}{57}, \frac{1}{57}, \frac{0}{57}, \frac{1}{57}]^T = [0.81, 0.16, 0.02, 0, 0.02]^T$. This prominent divergence between $\rho(C_2)$ and $\rho(data)$ strongly evidences against a random grouping, suggesting a unique pattern that distinguishes the samples in this cluster from the rest. Hence the pattern revealed by this group of sample may provide high value for the subtype classification based on the differential transcription of the corresponding ASM.

In order to measure the divergence between the subtype distribution of a cluster and that of the entire data, we define the *information score* of a cluster as below.

Let $m^1, m^2, \cdots, m^k$ denote the number of samples in the k groups, and let $m = \sum_{i=1}^{k} m^i$ denote the size of the dataset. The original subtype composition of the dataset is

$$\rho(data) = [q^1, q^2, \cdots, q^k]^T = [\frac{m^1}{m}, \frac{m^2}{m}, \cdots, \frac{m^k}{m}]^T. \tag{6.5}$$

For a cluster C_i, let $n_i^1, n_i^2, \cdots, n_i^k$ denote the number of cluster members from each subtype group, and let $n_i = \sum_{j=1}^{k} n_i^j$ denote the size of the cluster. Then the

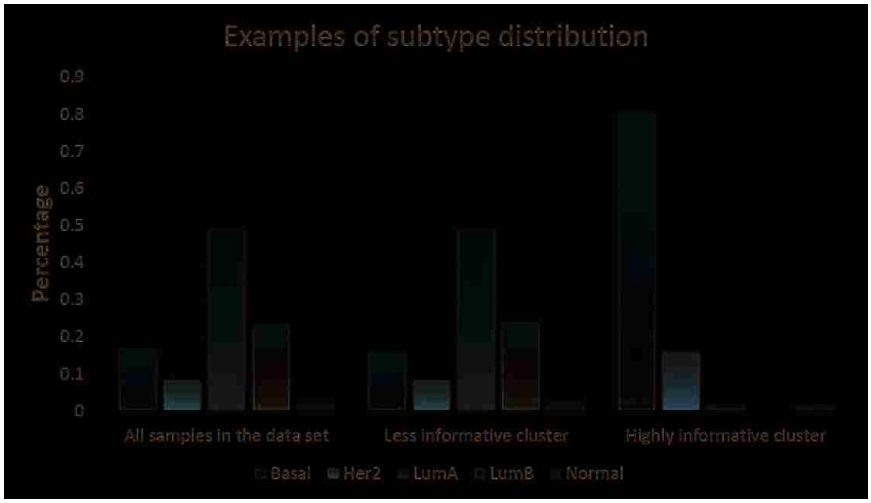


Figure 6.3: Examples of subtype distribution of a cluster. **Left:** subtype distribution of all samples in the data set. **Middle:** subtype distribution of a less informative cluster, a distribution with similar shape as the overall subtype distribution of the entire data set. **Right:** subtype distribution of a highly informative cluster — most samples that constitute this cluster come from Basal and Her2 hence the differential transcription captured in this alternative splicing module may reveal important information in classifying Basal and Her2 samples from the rest subtypes.

observed subtype distribution of C_i is

$$\rho(C_i) = [p_i^1, p_i^2, \cdots, p_i^k]^T = [\frac{n_i^1}{n_i}, \frac{n_i^2}{n_i}, \cdots, \frac{n_i^k}{n_i}]^T. \tag{6.6}$$

The original subtype composition of the entire dataset $\rho(data)$ is the maximum likelihood estimator for the null model, the cluster being drawn randomly from the population. The observed subtype distribution is the maximum likelihood estimator for the alternative model, the cluster being drawn from a distribution different from the composition of the population. The information score then compare the two

models by taking the log likelihood ratio,

$$
\begin{aligned}
IS(C_i) &= \log \frac{L(\text{Alternative model})}{\text{Null model}} && (6.7)\\
&= \log \frac{\prod_{j=1}^{k} (p_i^j)^{n_i^j}}{\prod_{j=1}^{k} (q^j)^{n_i^j}} && (6.8)\\
&= \sum_{j=1}^{k} n_i^j \cdot (\log p_i^j - \log q^j). && (6.9)
\end{aligned}
$$

Unlike a metric directly operated on the distribution vector, such as the Euclidean distance and the Jensen-Shannon divergence, the information score we have defined also takes into account the size of the cluster.

Larger information score reflects heavier deviation of the observed subtype distribution from the underlying subtype distribution and may help pick highly informative clusters.

6.4 Transcriptome analysis on TCGA breast cancer dataset

6.4.1 Differential transcription of breast cancer subtype groups

A total of 184,663 splice junctions have been detected by the RNA-seq reads of all BRCA samples, amongst which 161,896 (87.7%) are in existing transcriptome annotation. These splice junctions, together with the exonic segments reconstructed, have constituted a total of 6,745 ASMs, 4,237 (62.8%) of which contain novel junctions. The most common category of the ASMs observed in BRCA samples is alternative transcription start/termination, which 3,134 (46.5%) of all ASMs belong to. A large amount of ASMs (2,001 or 29.7%) have exhibited as combinations of two or more ASMs of different categories. The distribution of the basic ASM categories has been

summarized in Table 6.1.

We first tested all ASMs for consistent groupwise differential splicing, *i.e.*, for every ASM we tested the null hypothesis that the splicing isoforms of the ASM have the same proportion in all subtype groups. Under false discovery rate (FDR) ≤ 0.01 and minimum pairwise square root of JSD ≥ 0.1, we have identified a total of 1,206 ASMs, amongst which 847 (70.2%) contain novel splice junctions. Some genes many have multiple ASMs with significant differences. The 1,206 significantly differentiated ASMs have come from 1,118 genes. The distribution of categories of the significantly differentiated ASMs is similar as that of all ASMs, indicating that the category of ASM and the number of splicing isoforms have not caused obvious bias in the evaluation of statistical difference (Table 6.2).

Table 6.3 summarizes the number of significantly different ASMs between any pair of subtype groups. The normal-like samples exhibited the largest heterogeneity with samples in the other subtypes. Over 500 ASMs were identified as significantly different between normal-like subtype and any other subtype. The basal group has shown a large amount of different ASMs from HER2 and LumB, and LumA samples have shown similarity as LumB samples.

6.4.2 Cluster analysis of transcription profiles of alternative splicing events

The cluster analysis has been conducted on 1740 ASMs tested to exhibit differential transcription in the 728 tumor samples of TCGA dataset. A total of 143 clusters have been selected as potentially informative transcription patterns whose information scores are > 20 and whose sizes are ≥ 10. These clusters come from 119 ASMs of 69

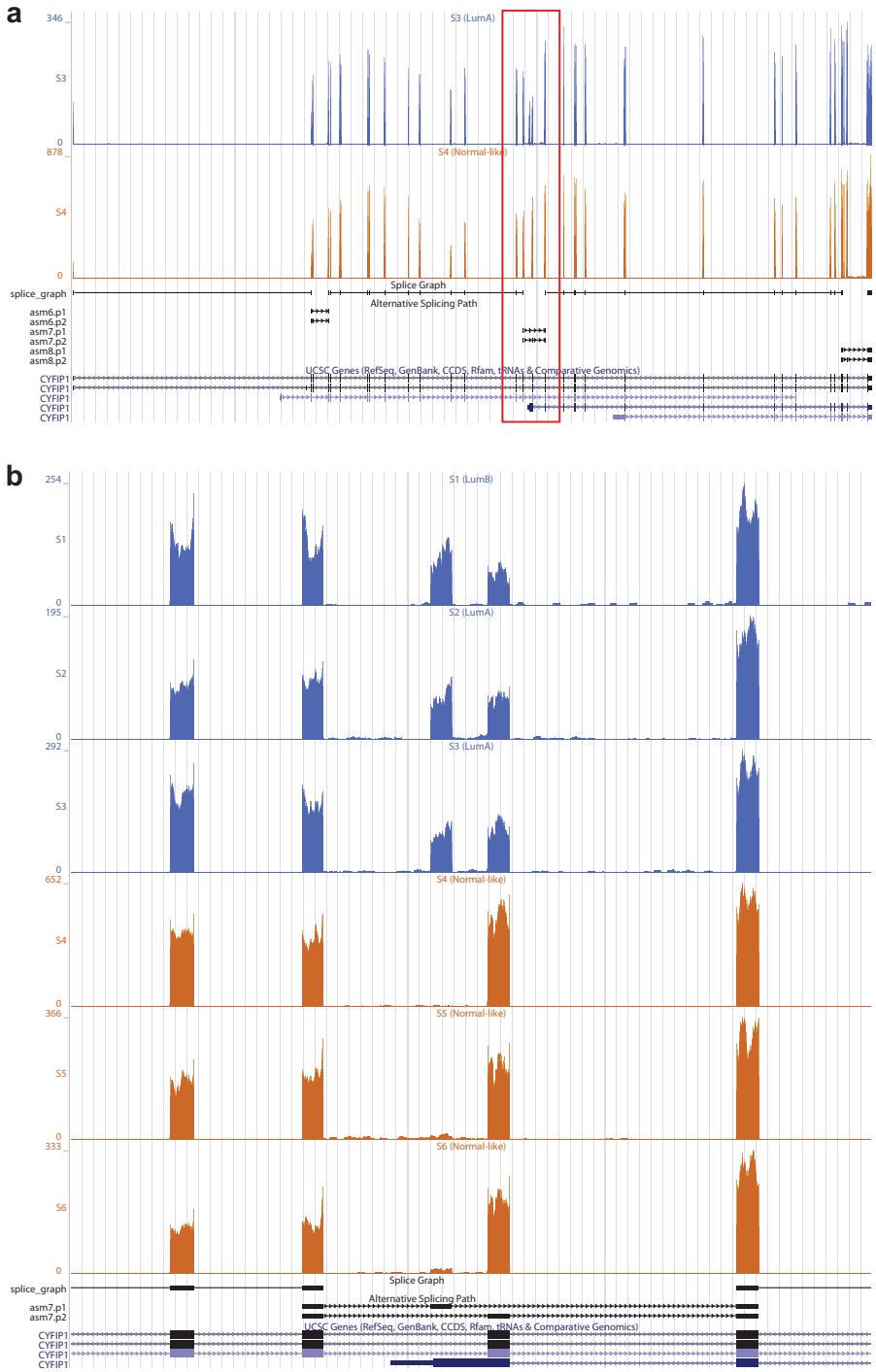


Figure 6.4: A novel alternative splicing event which is also differentially spliced in different subtypes within gene CYFIP1 in TCGA dataset. **(a)** The read coverage of gene CYFIP1. Three ASMs $asm6-8$ have been identified in this gene. **(b)** The ASM $asm7$ shows the mutual exclusive exons in this gene. The 3rd and 4th exons are contained in alternative transcription paths p_1 and p_2, respectively, but they are not included in the same path. The splice junctions and exon in p_1 is novel to the reference transcripts. The path p_1 has similar expression as p_2 in LumA and LumB samples, but has very low expression in normal-like samples.

genes, including 57 clusters from regions not in the current gene annotation.

In Figure 6.5 we show the expression of gene EVL in 6 randomly selected samples (2 LumA, 1 LumB, 2 Basal, 1 Her2) of TCGA breast cancer dataset. According to the overall gene read coverage, the LumA and LumB samples (tracks in blue) have higher expression level of this gene than the Basal and Her2 samples (tracks in orange). However the variance of the expression level may be prominent even for samples classified as the same subtype. For example, the two samples in LumA has read coverage of 2000 and 300 respectively. Besides the change of gene expression, the transcription of EVL differs among the samples in the first ASM of the gene, $chr14.asm521$, whose 4 alternative transcription paths reveal 4 different sets of exons the transcripts in EVL may start with. The alternative splicing event captured in this ASM is recognized as an event of alternative transcription start sites. The splice junction in path p_1 is novel connecting the 2nd and the 4th exon, as compared to the transcript annotations. The 2nd exon, the alternative transcription start site identified by p_1 and p_3, is mainly expressed in LumA and LumB samples (colored blue), but tends to have very low expression in Basal and Her2 groups (colored orange). Figure 6.6a plots the distribution of transcription profiles of all tumor samples on this gene, coordinated by the log read coverage of transcription paths p_2, p_3 and p_4 in each sample. The colors distinguish samples from different subtype. Consistent with the samples selected for the coverage plot, most LumA and LumB samples tend to have larger expression of p_3 as compared to other subtypes.

Table 6.4 shows the 6 clusters resulting from the clustering of transcription profiles on this gene. The clusters 1 is highly informative with an information score of 74.22.

Table 6.1: The distribution of ASM categories in BRCA dataset. ES: exon-skipping; ME: mutual exclusive; IR: intron retention; ASS: alternative splice sites; ATS/ATT: alternative transcription start/termination; Mixed: a combination of two or more ASM categories.

ASM Category	ES	ME	IR	ASS	ATS/ATT	Mixed
Number of ASMs	831	413	102	264	3134	2001

Table 6.2: The distribution of ASM categories in BRCA dataset, for the ASMs with significant differences in transcription. ES: exon-skipping; ME: mutual exclusive; IR: intron retention; ASS: alternative splice sites; ATS/ATT: alternative transcription start/termination; Mixed: a combination of two or more ASM categories.

ASM Category	ES	ME	IR	ASS	ATS/ATT	Mixed
Number of ASMs	113	10	5	28	568	482

Table 6.3: The distribution of the number of significantly different ASMs between two subtypes.

	HER2	LumA	LumB	Normal-like
Basal	540	173	543	850
HER2		355	305	758
LumA			97	543
LumB				821

Table 6.4: Clustering of samples on transcription of EVL in TCGA dataset. The clustering of samples on the transcription of the alternative transcription start sites in gene EVL shows 6 clusters. The clusters 1, 2 and 4 are considered to have higher information score. The cluster 1, in particular, consists of mostly Basal samples.

Cluster	Score	Size	Basal	Her2	LumA	LumB	Normal
1-blue	74.22	57	46(81%)	9(16%)	1(2%)	0(0%)	1(2%)
2-cyan	17.48	23	10(43%)	8(35%)	1(4%)	3(13%)	1(4%)
3-purple	2.91	2	0(0%)	0(0%)	0(0%)	2(100%)	0(0%)
4-red	14.88	601	61(10%)	38(6%)	340(57%)	152(25%)	10(2%)
5-green	2.71	40	5(13%)	5(13%)	16(40%)	11(28%)	3(7%)
6-black	2.55	5	1(20%)	0(0%)	1(20%)	2(40%)	1(20%)

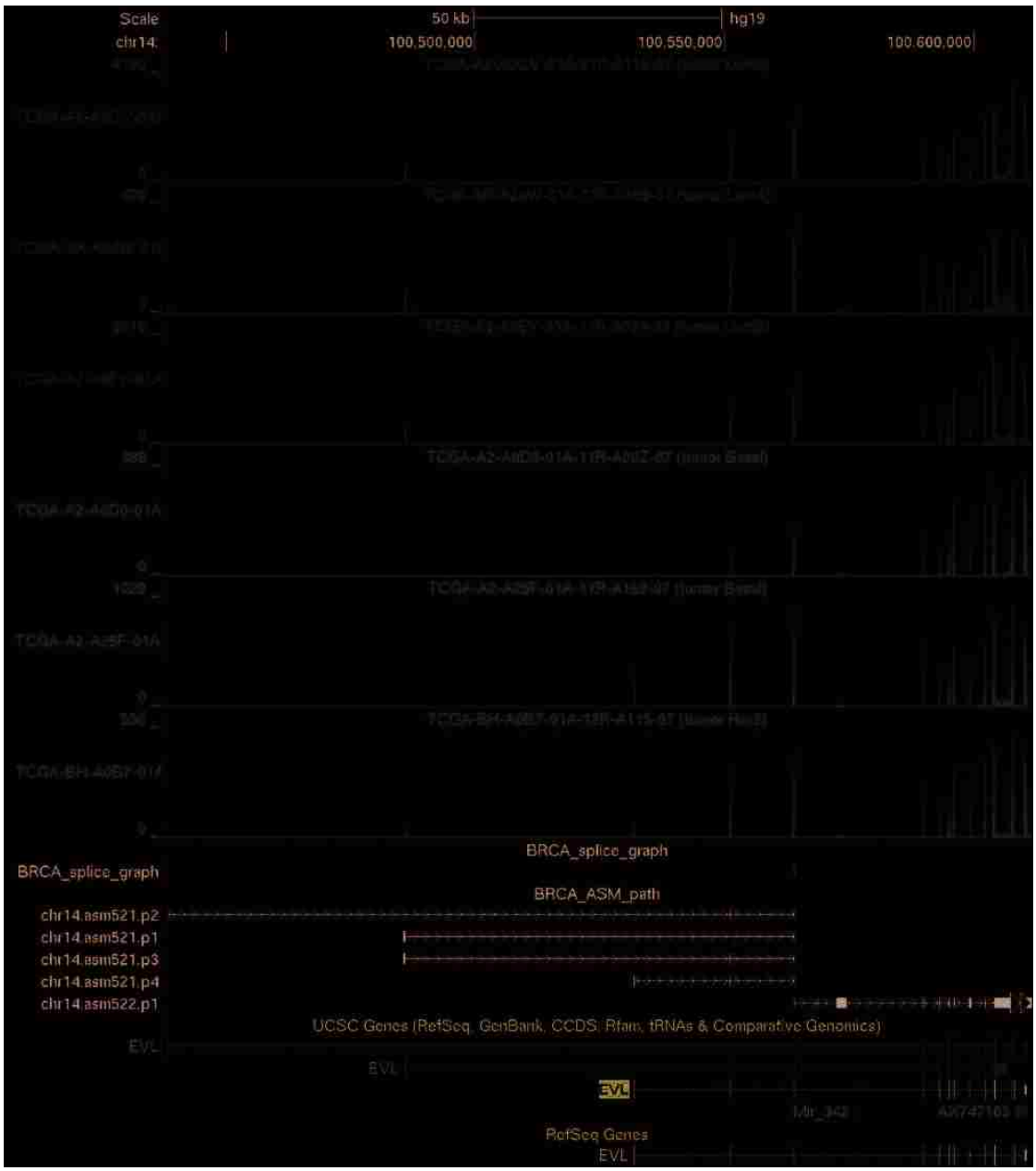
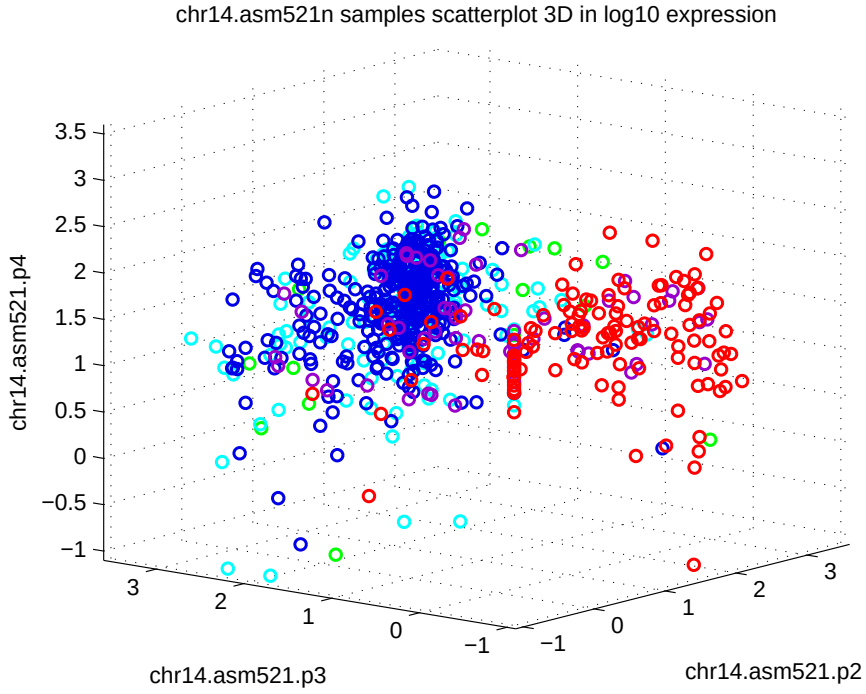
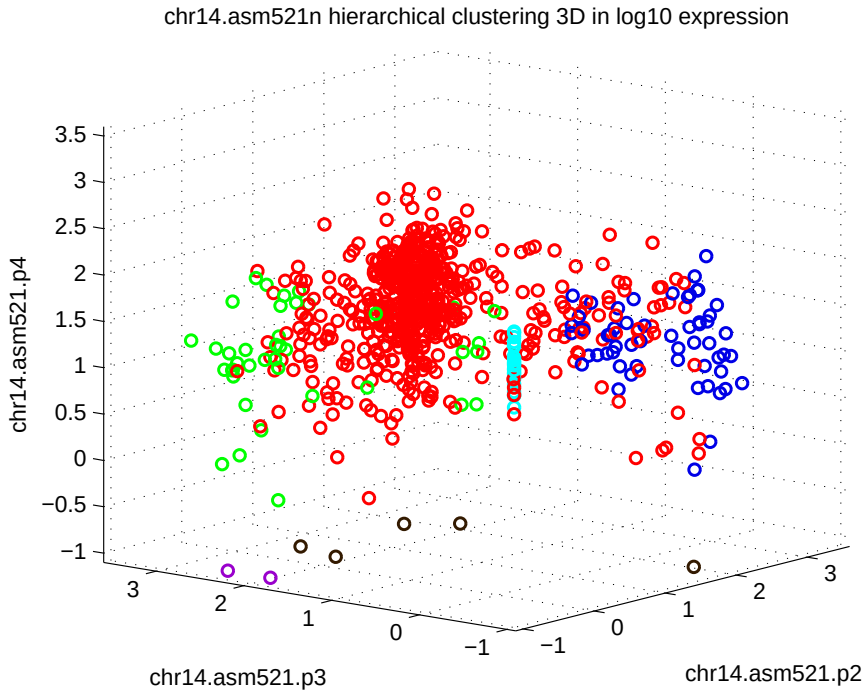


Figure 6.5: Expression of gene EVL in TCGA dataset. The ASM $chr14.asm521$ shows the alternative transcription start sites in this gene. Four alternative transcription paths $p_1 - p_4$ correspond to transcripts starting with 4 different sets of exons in the gene. The splice junction in p_1 is novel to the reference transcripts.



(a)



(b)

Figure 6.6: The 3-D scatter plot of samples in the transcription of gene EVL in TCGA dataset. **(a)** The scatter plot color coded by subtypes: red-Basal, purple-Her2, blue-LumA, cyan-LumB, green-Normal. **(b)** The scatter plot color coded by clusters. The three dimensions correspond to the read coverage of transcription paths p_2, p_3 and p_4, in logarithm scale.

This cluster consists of 57 samples, 81% of which come from Basal group. The samples in this cluster all have very low (read coverage ¡1) expression on p_3, as shown in the blue points of Figure 6.6b.

Copyright © Yin Hu, 2013.

Chapter 7 Other related work in RNA-seq-based transcriptome analyses

Although the application of next-generation sequencing or high-throughput sequencing technologies on transcriptome analysis has not been long, just since year 2008 [Marguerat et al., 2008, Wang et al., 2009], there have been abundant research work that utilizes the deep sequencing coverage on a transcriptome of interest for insights into the linkage from genotype to phenotype of various species.

This chapter summarize the typical workflows of RNA-seq analysis and some of the central challenges that the computational methodologies are developed to address (Figure 7.1). The diagram in Figure 7.2 further summarizes the data flows for the analyzing pipelines, from RNA-seq reads to each computational problem.

7.1 Reference-guided gene level expression analysis

A gene is the basic unit of the hereditary information stored in a cell's DNA and is passed from parents to the offsprings. It is considered that the change of expression level of a gene may result in the change of density of functional products encoded by the gene, such as the RNA or protein. The altered RNA density and protein expression may further affect the function of the cell. Therefore, the detection of differential gene expression constitutes an essential tool for the understanding of cell differentiation and disease.

In microarrays, the gene expression level is read as a continuous value, and statistical tests may be formed to evaluate the equivalence between the expression levels

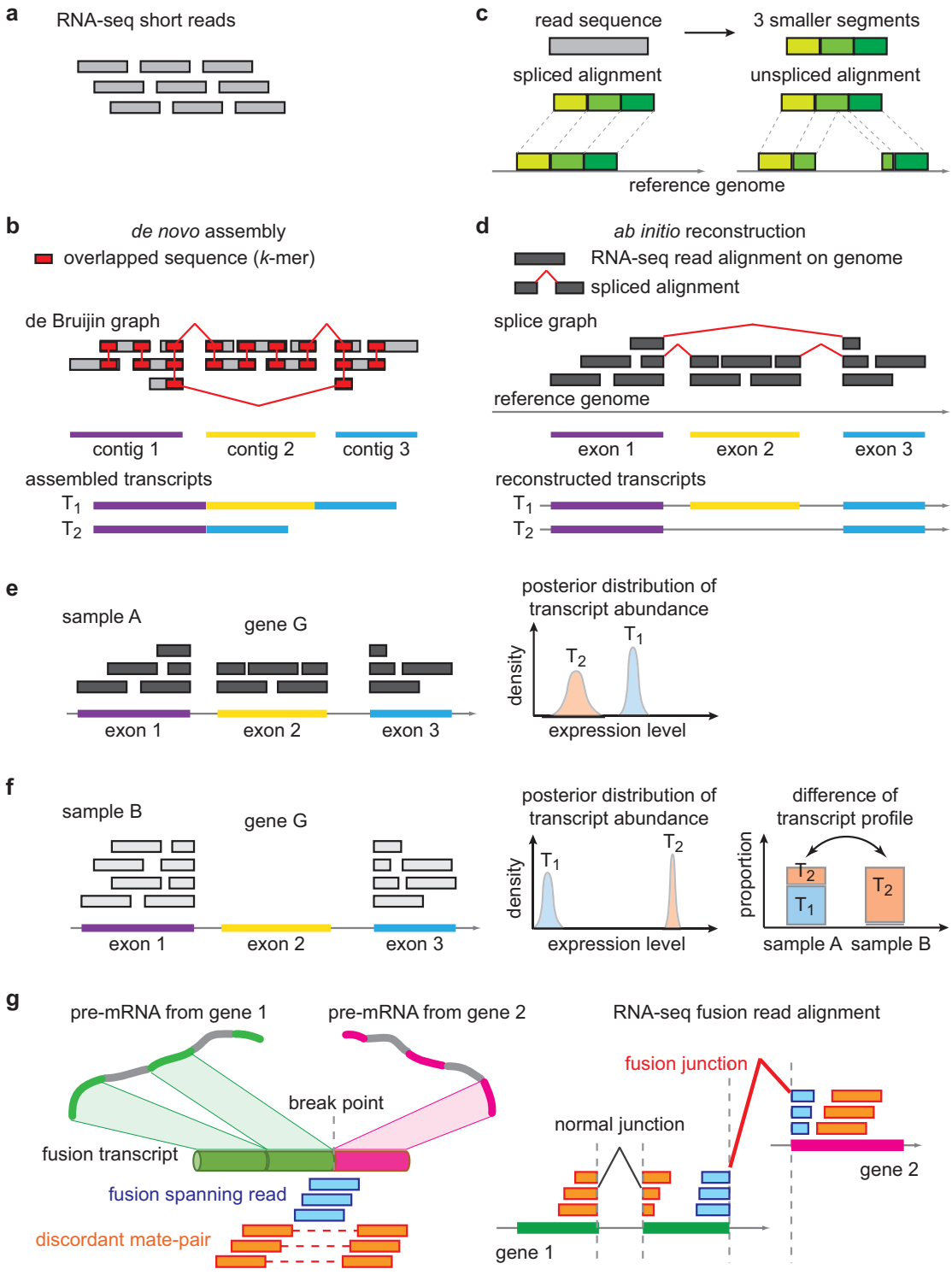


Figure 7.1: Typical workflows and computational challenges in transcriptome studies using RNA-seq.**(a)** The RNA-seq short reads. **(b)** *De novo* transcriptome assembly. **(c)** RNA-seq short read alignment to the reference genome. **(d)** *Ab initio* transcriptome reconstruction. **(e)** Transcript abundance estimation. **(f)** Differential transcription analysis. **(g)** Gene fusion detection.

of a gene in different conditions. Fold change, the ratio of the expression level in one condition against that in another condition, is often calculated to measure the practical magnitude of the difference. The RNA-seq-based differential gene expression analysis extends from that for the microarray data, except for that the expression measurement is now discrete as the read count. With a reference genome or transcriptome, the RNA-seq reads may be mapped to reference by matching the (short) sequences of the reads to the (long) sequence of the reference. The number of RNA-seq reads falling into each gene is counted. These counts are treated as the observed expression on the genes and are then modeled by a probability distribution, typically the negative binomial distribution. Finally, gene expression in multiple samples is jointly considered in the model to estimate the mean expression and the variance, allowing statistical tests for the equality of gene expression between sample groups.

Biologically, this analysis does not explore the transcriptome expression and may not directly point to protein isoforms. Moreover, the count-based expression measurement without considering the transcripts may be biased, because the transcripts in the gene typically have different lengths. Computationally, the mapping of the read sequences to the reference sequences is not trivial. Many reads are sampled at the boundaries of two exons in the mRNA transcripts. These reads cannot be mapped to the reference genome as a entirety. Even when mapping to a reference transcriptome, perfect match may not exist due to individual modifications such as SNPs, insertions and deletions.

7.2 Mapping RNA-seq reads to a reference genome

For species with a reference genome available and well annotated, like the human, it is often preferred to map the RNA-seq short reads to the reference genome first. This procedure tries to find matches on the reference genome for the sequences of the sampled short reads. Importantly, gaps are allowed in the read alignments for up to 50,000bp typically. A read whose sequence can directly match a piece of the reference genome is mapped as an entirety and has an *unspliced alignment*. A read whose sequence should be split into segments and matched to different places of the reference genome separately has an *spliced alignment*. The spliced RNA-seq read alignments then suggest the splice junctions and introns on the genome, including those not known in existing annotation database.

The challenge for this spliced read mapping is the numerous possible splice positions in a read sequence and possible locations to to map on the reference genome. For example, consider only the case in which one gap is allowed in a read alignment, and a 100bp read will have 99 possible splice positions in theory. Each segment will then be compared to the $> 3 \times 10^9$ nucleotides of the reference genome for possible matches. Furthermore, because the genome sequence consists of only 4 characters, a segment too short may have too many matches (also called hits) on the genome. The computational burden may be further increased by repetitive sequences on the genome and mismatches due to individual insertions/deletions to the reference as well as sequencing errors.

One commonly used approach is to divide a read sequence into several smaller

segments, or k-mers. The length of the segments, k, is usually set to be ≥ 8, in order to keep the complexity of the segment sequence and to reduce the number of possible hits on the reference genome. Sequence matching on the reference genome is tried for every segment. A segment that cannot be mapped is then expected to contain a splice point. It will be split to locate the potential splice junction, by searching regions surrounding the mapped positions of the segments next to the unmapped segment.

The read mapping approaches can be divided into two categories, the unspliced aligners and the spliced aligners [Garber et al., 2011]. The unspliced aligners are typically used to align reads to a reference transcriptome, including MAQ [Li et al., 2008a], SHRiMP [Rumble et al., 2009], ELAND [Cox, 2007], Novoalign [Hercus], Stampy [Lunter and Goodson, 2011], Bowtie [Langmead et al., 2009], BWA [Li and Durbin, 2009], Bowtie 2 [Langmead and Salzberg, 2012] and SNAP [Zaharia et al., 2011]. Reads are mapped to the reference as an entirety, possibly allowing several bases of mismatch. For example, SHRiMP [Rumble et al., 2009] and Stampy [Lunter and Goodson, 2011] take the seed-extension strategy. They first find matches in the reference for short subsequences of the reads, then apply on candidate regions around each seed more sensitive algorithms such as the Smith-Waterman algorithm or statistical models and lastly extend the seeds to full alignments. Alternatively, Bowtie [Langmead et al., 2009] and BWA [Li and Durbin, 2009] use the Burrows-Wheeler transformation to compress the reference genome and provide higher efficiency.

However, unspliced aligners can only be used to map the RNA-seq reads onto

known exons and known transcripts. Novel splices and novel transcript isoforms will then be missed. Therefore, the spliced aligners are often favored for transcriptome analyses, including QPALMA [Bona et al., 2008], TopHat [Trapnell et al., 2009b], SpliceMap [Au et al., 2010], MapSplice [Wang et al., 2010a], GSNAP [Wu and Nacu, 2010] and STAR [Dobin et al., 2013]. For example, utilizing a training set of positive controls, QPALMA [Bona et al., 2008] developed a machine learning algorithm to help predict splice junctions. Representing the exon-first approaches, TopHat [Trapnell et al., 2009b] first maps full, unspliced reads to the reference genome and reconstructs candidate exons. Unmapped reads are then divided into smaller segments and mapped to candidate splice junctions constructed from pairing candidate exons. SpliceMap [Au et al., 2010] uses the bases flanking the splice site to help locate potential splice sites. MapSplice [Wang et al., 2010a] divides reads into short segments (k-mers) and maps the segments to the reference genome using unspliced aligners like Bowtie. Splice sites will then be searched around mapped segments if their adjacent segments cannot be mapped. In principal, these approaches may find spliced read alignments separated by gaps from several base pairs up to hundreds of thousands of base pairs, or even read alignments spanning two chromosomes (such as the fusion reads detected by MapSplice). These spliced aligners then give great flexibility to transcriptome analyses by picturing the genome-wide splice graph purely following the data, allowing the study of subject-specific splices not cataloged by existing database or even fusion transcripts. In practice these procedures may be confounded by short read length, insertions/deletions, repetitive sequences, pseudogenes and sequencing errors.

7.3 Gene fusion discovery

In addition to normal transcriptome analyses as described above, **gene fusion detection** is another central application of RNA-seq, extremely useful for cancer transcriptome analyses. Along with alternative splicing, fusion transcripts in cancer cells may serve as prominent biomarkers that benefit tumor diagnosis, prognosis and treatment.

Gene fusion is the phenomenon that sequences of two separate genes are merged together into a hybrid gene, leading to disfunction RNAs but sometimes proteins with abnormal functions that may result in diseases like cancer. This may be caused by structural abberations such as chromosome translocation, the switch and rearrangement of parts between two chromosomes, deletion, the loss of a part of a chromosome, or chromosomal inversion, a part of a chromosome being reversed end to end.

In Figure 7.1f, for example, the left two exons of gene 1 (green) and the rightmost exon of gene 2 (pink) are concatenated into a fusion transcript that contains sequence from both genes. When sequencing the transcriptome, RNA-seq reads will cover the fusion break point if they span the green exon and the pink exon simultaneously. These reads are called the fusion spanning reads (blue reads). If paired-end sequencing protocols are applied, read pairs sequenced from transcript fragments that contain the fusion break point may have one end from gene 1 and the other end from gene 2. These reads are called the discordant mate-pairs (orange read pairs).

When mapping the RNA-seq reads to the reference genome, the fusion spanning reads should have no normal alignments, alignment either in an entirety or with proper segments separated by gaps of a reasonable length (in the region of a gene). Instead,

these reads should be mapped as segments separated by extraordinarily long gaps or on discordant strands or even on different chromosomes. The locations where the fusion spanning reads are split indicate the fusion junctions. In addition, discordant mate-pairs may have their end reads mapped around a fusion junction, one on each side. These discordant mate-pairs also suggest the existence of the fusion junction connecting the two genes.

In practice, the limited read length often results in a large set of fusion junction candidates (hundreds or thousands). Many false fusion alignments may be introduced due to repetitive sequences on the genome and mismatches in the read sequences. Because most fusion transcripts are not highly abundance, the classification of real fusion transcripts is difficult and often has a low specificity.

Related to the RNA-seq read mapping, there are also a handful of methods for the discovery of gene fusion. In cancer, alternative splicing and gene fusion events are commonly found in the mRNA transcriptome [Maher et al., 2009, Berger et al., 2010]. Gene fusion events were initially found in non-epithelial cancers, such as leukemia and lymphomas. Recent research using RNA-seq data also demonstrated gene fusion events in common epithelial cancers accounting for 90% of cancer related deaths [Maher et al., 2009]. Therefore, cancer specific splicing and fusion events are promising biomarkers and targets for diagnostic, prognostic and treatment purposes. Traditionally, paired-end reads have been applied to detect structural variations in DNA [Medvedev and Brudno, 2008, Lee et al., 2008, Medvedev et al., 2009, Hormozdiari et al., 2009, Maher et al., 2009]. The end reads are individually mapped and the pairing information is used to infer the regions where breakpoint may locate. The

candidate regions of breakpoints then suggest potential genes that fuse together, but the actual sequences of fused transcripts remain unknown.

Using RNA-seq reads, especially reads of 75bp or longer, the precise identification of actual fusion sites becomes possible. For example, extensions to spliced aligner for normal splices, such as MapSplice [Wang et al., 2010a] and TopHat-Fusion [Kim and Salzberg, 2011], have been developed to discover fusion junctions, abnormal splice junctions that have extraordinarily long intron size (*e.g.* ¿50,000) or that have splice sites on discordant strands or that have splice sites on different chromosomes. Similar approaches taking this mapping-based strategy include SplitSeek [Ameur et al., 2010], ShortFuse [Kinsella et al., 2011] and FusionSeq [Sboner et al., 2010]. Another strategy, represented by Trans-ABySS [Robertson et al., 2010], relies on the *de novo* assembly of transcripts and then map the assembled transcripts for discordantly mapped ones across fusion points. Other computational approaches for the discovery of gene fusion events and structural variations include SnowShoes-FTD [Asmann et al., 2011], FusionMap [Ge et al., 2011], FusionHunter [Li et al., 2011b], Comrad [McPherson et al., 2011b], deFuse [McPherson et al., 2011a], nFuse [McPherson et al., 2012], ChimeraScan [Iyer et al., 2011] and Dissect [Yorukoglu et al., 2012]. The performance of these methods very much depends on the software implementation and parameter setting, and different approaches may have significant disagreement even when applied on the same dataset.

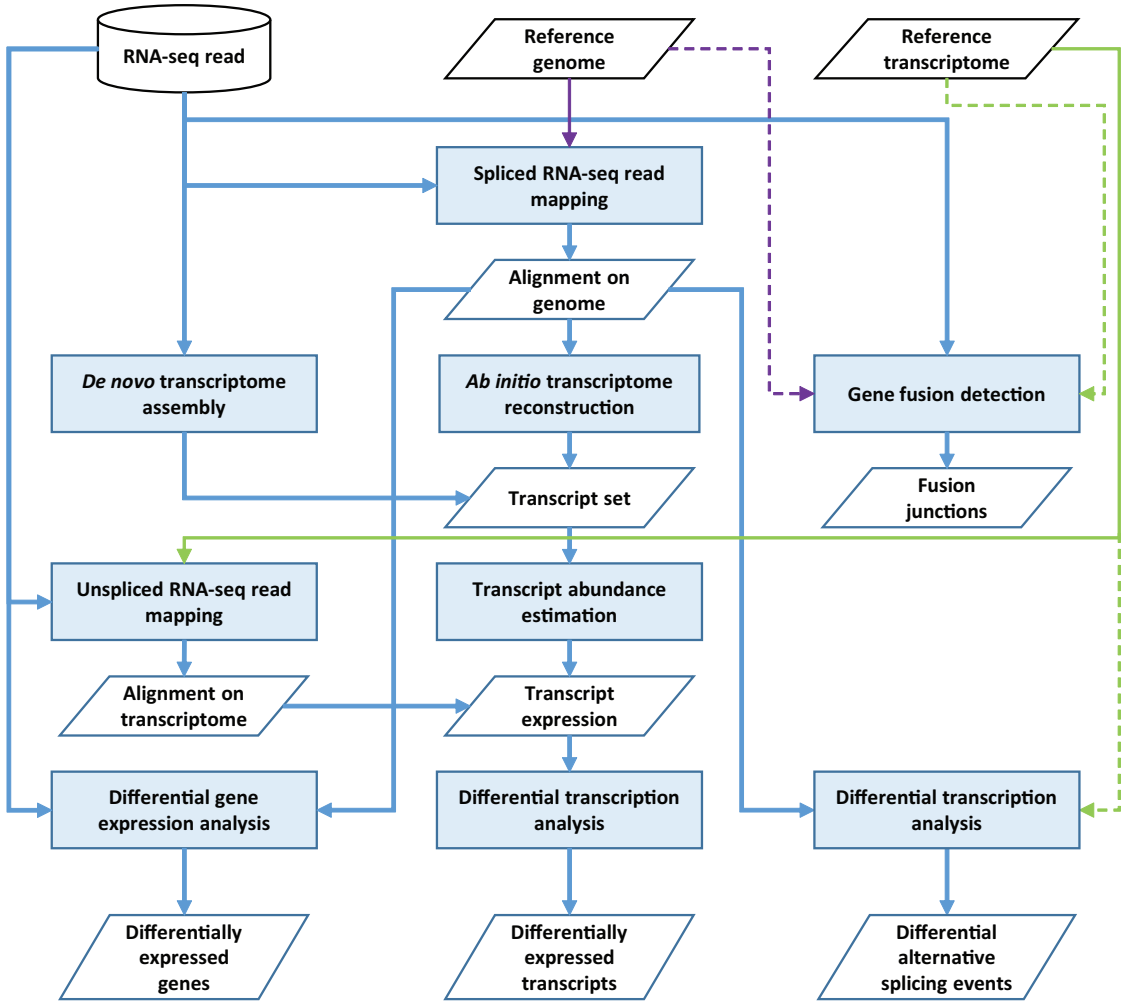


Figure 7.2: The primary computational problems and typical workflows in transcriptome studies using RNA-seq technologies. The blue arrows show the data flows. The purple arrows point to processes that require a reference genome and the green for requiring a reference transcriptome. The dashed arrow indicates that the reference is not necessarily required by all methods. Some issues are omitted in the chart but have the same importance, such as the normalization across samples before differential analyses and the correction of sampling biases.

7.4 Reference-free *De novo* genome/transcriptome assembly

Besides the enormous amount of reads that may be collected to suggest the gene expression levels, the greatest power of RNA-seq technologies is their ability to discover novel transcripts, transcripts that have not been cataloged in existing knowledge databases. The *de novo* transcriptome assembly problem has emerged as the assembly of transcripts in the original sample solely based on the sequences of the short reads.

Because the transcriptome is deeply sampled, it is expected that most bases of a moderately abundant transcript are sampled and sequenced by multiple reads. Reads are typically divided into smaller seeds (k-mers, where k denotes the length of the seed) for more efficient identification of overlaps. Two reads may be considers as coming from a same transcript if their sequences share significant overlaps, *e.g.*, a k-mer. The read sequences may then be summarized using a compact representation called the de Bruijin graph. Seeds are gradually extended into contigs, and reads overlap with multiple contigs suggest the splice connection between contigs. The de Bruijin graphs may then be traversed to assemble contigs into transcript sequences. (Figure 7.1b)

Due to the limited read length and hence small k-mers, the difficulty of the assembly mainly lies in the numerous possibilities of overlapped read sequences. Small sequence modifications such as SNPs and insertions/deletions, together with sequencing errors, will further confound the assembly by introducing extra paths to the de Bruijin graph. As a result, the generated de Bruijin graph often has a size and

complexity infeasible for transcript assemble. The length and the occurrence of the overlaps are then common criteria to filter the graph, at a risk of overlooking transcripts with relatively low expression. In practice, *de novo* assemblers often require considerable computational power and resources (tens or even hundreds of RAM, running for days or weeks), and may not provide satisfying sensitivity and specificity of the assembled transcript set.

In order to enable the short-read sequencing on transcripts at random positions, the mRNA molecules will be randomly fragmented into pieces before library preparation. Transcript fragments of proper size (typically 200bp to 500bp, depending on the desired insert size) will be selected as the cDNA library and sequenced at one end or both ends if sampled. Thus the observed RNA-seq reads are only local pieces of the original transcripts. A group of *de novo* assembly approaches have been proposed for the reconstruction of the mRNA transcriptome on the basis of only the sequenced reads.

Short-read assemblers were first developed for the assembly of genomes, including Velvet [Zerbino and Birney, 2008], ALLPATHS [Butler et al., 2008] and ABySS [Simpson et al., 2009]. For example, Velvet [Zerbino and Birney, 2008] uses the de Bruijn graphs to represent the overlapped short sequences (captured by k-nucleotides or k-mers) among the short reads (25-50bp) in the sample, removes errors such as erroneous cycles and connections in the graph, then produces high quality unique contigs. This algorithm set may further use paired-end read and long read information to help resolve repeats and scaffold contigs in large complex genomes [Zerbino et al., 2009].

Direct application of these *de novo* genome assembly approaches on transcriptome

assembly is not favorable. In RNA-seq, the dynamic range of measured transcription expression can be very large, and a small portion of highly abundant transcripts may dominate the majority of the sampled reads [Wang et al., 2009]. The large variation in local read densities may cause problems to parameter setting in filters of noises and repetitive regions, which may lead to falsely throwing transcripts with very high abundance or very low abundance. The assembly of transcriptome is further complicated by the large amount of shared sequences among transcripts from the same gene. In addition, many RNA-seq protocols are strand-specific, allowing the identification of the transcription direction from the reads. For protocols that are not strand-specific, the strand of transcription may also be inferred by the dinucleotide sequences in the intron flanking the donor and acceptor sites. Then the strand information should also be taken into account for transcript assembly to enable the knowledge of the actual mRNA sequence and to help separate genes with overlapped exonic sequences. [Martin and Wang, 2011]

Most *de novo* transcriptome assemblers still adopt the compact representation of the de Bruijn graph. One strategy is to assemble the reads multiple times using an algorithm such as Velvet, ALLPATHS and ABySS with varying stringencies. This will allow the discovery of transcripts with a broad range of expression levels. The final set of transcripts in the sample will then be curated by merging transcripts resulting from individual runs and removing redundancy. Approaches taking this strategy include Rnnotator [Martin et al., 2010], Multiple-k [Surget-Groba and Montoya-Burgos, 2010] and Trans-ABySS [Robertson et al., 2010]. Other assemblers, such as Trinity [Grabherr et al., 2011] and Oases [Schulz et al., 2012], directly assem-

ble transcripts by traversing the de Bruijn graphs. For example, Trinity first uses a greedy path-searching algorithm in the k-mer graph to recover a collection of linear contigs that best represent the alternative variants sharing k-mers. It then builds de Bruijn graphs for pools of contigs overlapped by $k1$-mers or connected by junctions, followed by trimming spurious graph edges. The graphs will lastly be compacted and consolidated with original reads, and graph paths will be extracted for the sequences of transcript isoforms.

The *de novo* assembly approaches are greatly useful when a reference genome is not available or when individual modifications to the reference genome is significant. However, due to the limited length of the short reads, the overlaps between each read on which the assembly relies are often short and bring high ambiguity, leading to considerable computational requirements and unsatisfied sensitivity and specificity.

7.4.1 *Ab initio* transcript reconstruction based on RNA-seq alignments

Provided the RNA-seq read alignment, the *ab initio* transcript reconstruction approaches are often preferred for a higher accuracy and efficiency than the *de novo* strategy.

Because the sequences of RNA-seq reads are those kept in the mRNA transcripts, the genomic coordinates which have RNA-seq reads aligned to can help recover the exonic sequences on the genome. Nucleotides may be considered forming an exon if they are contained in a same read or a same mate-pair, or if they locate close to each other on the genome (*e.g.*, several bases apart) with no spliced alignments in between. The splice junctions will indicate how the exons should be concatenated

during splicing. Then a graph can usually be constructed to picture the connectivity in a potential gene of all the reconstructed exons or of all the read alignments. Transcripts may further be identified by traversing the graph for confident graph paths.

In Figure 7.1c, for example, the read alignments have suggested three exons connected by three splice junctions. Two transcript isoforms can be reconstructed from the splice graph, one retaining the middle exon and one skipping. Because the read alignments can capture exons and junctions both from existing knowledge or novel, the reconstructed transcripts may discover uncataloged isoforms beyond transcriptome annotation databases.

In real-world RNA-seq data, read alignments may contain a large amount of spurious splice junctions. The identification of exons may be also complicated by insufficient sampling, intron noises and read mapping issues. As a result, the splice graph may be too complex or disconnected. Heuristics such as maximum parsimony or abundance-based regularization are what computational approaches usually rely on, for the identification of a most probable and compact set of transcripts.

Provided with the RNA-seq read alignments, a more favorable pipeline to profile the transcriptome with short reads is the reconstruction of transcripts from the read alignments. This is typically denoted as the *ab initio* reconstruction.

Because the reads were sampled from the mRNA transcripts, genomic loci that are covered by read alignments should then correspond to exonic sequences that eventually constitute the mRNA molecule. The reads with spliced alignments were sampled at the boundary of two consecutive exons in an mRNA transcript. The gaps in the

spliced alignments then reveal the splice junctions on the genome, indicating the exon boundaries and how the exons should be connected (Figure 7.3a). Thereafter, a splice graph [Heber et al., 2002, Sammeth, 2009, Singh et al., 2011, Hu et al., 2012] may be constructed consisting all inferred exons (typically graph vertices) and splice junctions (typically graph edges) (Figure 7.3b). The edges may further have directions that show the transcription strand, making the splice graph acyclic. Computationally, a candidate transcript then corresponds to a path in the splice graph. The final set of transcripts will be derived after filtering using transcript abundance or biological knowledge and possible compacting using long reads or paired-end reads [Feng et al., 2011, Li et al., 2011a, Guttman et al., 2010]. Alternatively, Cufflinks [Trapnell et al., 2010] constructs a read-level overlap graph to represent the transcription compatibility among all read alignments (Figure 7.3c). Alignments (corresponding to vertices in the overlap graph) that may possibly come from the same transcript are connected by edges. The set of transcripts is then derived as the minimum path cover in the graph.

The performance of the reconstruction highly depends on the complexity of the gene model and the quality of the read alignments. If read alignments miss exonic sequences or splice junctions, the derived transcript set will probably lose transcript isoforms or have exons discontinued in consequence. More often observed are complex regions where read alignments have suggested too many spurious splice junction, leading to too many candidate paths (transcripts) in the splice graph. The primary difficulty in resolving these obstacles is the ambiguity in linking variants of different alternative splicing events into full-length transcripts. Even the paired-end reads may

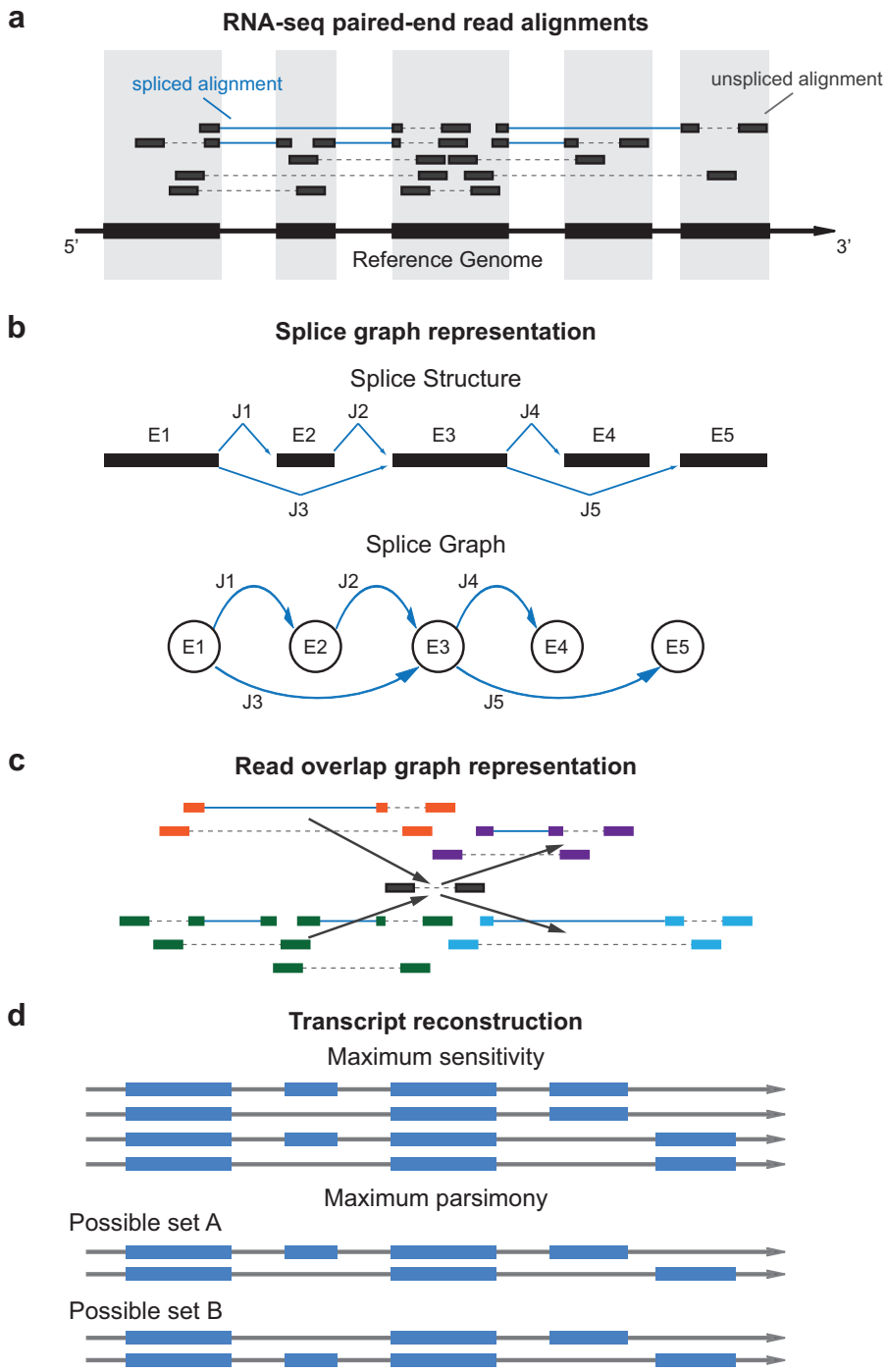


Figure 7.3: An illustration of the *ab initio* transcript reconstruction pipelines. **(a)** Alignments of paired-end RNA-seq short read to the reference genome. **(b)** The splice graph representation depicting the connectivity of exons via splice junctions. Possible transcripts correspond to valid paths in the graph. **(c)** The read overlap representation depicting the compatibility of read alignments. Possible transcripts correspond to a path cover in the graph. **(d)** The heuristics in transcript set selection. The maximum sensitivity takes all possible transcripts as condidate for future filtering. while the maximum parsimony takes the minimum set of transcripts capable of explaining the splice variants from the read alignments.

provide only local pieces of evidence up to several hundreds of base pairs (typically ¡500bp), usually insufficient to span all alternative splicing events in a gene thus unable to uniquely identify the complete transcript composition [Garber et al., 2011]. In such case, heuristics are what the existing *ab initio* methods rely on (Figure 7.3d). For example, on the basis of maximum parsimony, Cufflinks [Trapnell et al., 2010] will choose the minimum set of transcripts that can explain the observed paired-end read alignments. Following maximum sensitivity, Scripture [Guttman et al., 2010] will keep all putative isoforms, subject to later biological filtering. Other methods, such as IsoLasso [Li et al., 2011a], apply L1-regularization (known as Lasso) to reinforce transcript set shrinkage by favoring candidates with higher estimated abundance. Each of these transcript-level heuristics reflects a general sense about what true transcript isoforms could look like. For example, the philosophy behind maximum parsimony is that the most concise set of transcripts necessary to explain the data tends to have sufficient sensitivity and high specificity, and that behind regularization is that transcripts with very low expression are likely the artifacts due to sampling ambiguity.

7.4.2 Transcript abundance estimation

The number of RNA-seq reads falling in a gene may be collected as a measure for the gene's expression level. However, transcripts in a gene typically share sequences. It is not trivial to identify the original transcript for reads mapped to common exons where transcripts overlap. As a result, the transcripts' expression level may not be derived by counting the RNA-seq read alignments. The *transcript abundance estimation* is then a procedure that estimates the expression level of the transcripts

based on the read alignments. The transcript abundance estimation approaches have been discussed in Section 5.2.

For example, in Figure 7.1d, a total of 11 reads are mapped to gene G. Because both transcripts T_1 and T_2 contain exon 1 (purple) and exon 3 (blue), the reads mapped to these two exons may come from either T_1 or T_2. Reads mapped to exon 2 (yellow), on the other hand, can only come from T_1. Based on reads mapped to sequences unique to a subset of transcripts and leveraging the reads mapped to the shared sequences, the transcripts in the gene G may have expression estimates with plotted posterior densities. The modes of the two posterior densities indicate that, given the observed read alignments, the most probable abundance of T_1 is approximately 2 times that of T_2. The width of the density shapes then suggest that the inference about T_1 may be more precise than T_2 with a less variance. Point estimates may also be drawn with similar observations.

Although a handful of probabilistic models have been developed to infer the transcript abundance from the read alignments, the inference may be problematic under certain circumstances. For example, the accuracy and reliability of the inference tend to decrease quickly when the number of transcripts in a gene increases. The main reason is that transcripts may share a significant amount of sequences, making determining the origin of the reads highly ambiguous. Various types of sampling biases in RNA-seq may also affect the inference. For example, more reads may be attracted by the GC-content regions (regions rich of nucleotides 'G' and 'C'), and the exons at the start or end of a transcript may get less reads. The non-uniform sampling distribution may break the assumptions of an inference procedure and bias the result.

7.4.3 Differential transcription analysis

Transcriptome is known to vary in response to cellular differentiation and environmental change. Transcriptomes of cells from different tissues or different conditions may differ in the diversity of the transcript set as well as the abundance of the transcripts. Accordingly, the RNA-seq reads sampled from the transcriptomes will change. The transcript abundance estimation approaches have been discussed in Section 4.2.

In Figure 7.1e, for example, a group of RNA-seq reads from sample B are also mapped to the same gene as in Figure 7.1d. However, the read distribution on this gene is different in sample A and sample B, on exon 2 in particular. The read distribution in sample B suggests a higher expression of transcript T_2 but very low expression of T_1. Comparing the two samples, the dominant transcript switches from T_1 in sample A to T_2 in sample B. The divergence between the profiles of transcript abundance in the two samples may be further described quantitatively by comparing the transcript proportions or their absolute expression levels. Statistical significance may also be drawn for the observed divergence in transcription.

Computational approaches for analyzing differential transcription often take one of three strategies. The first strategy compares across samples the read distribution on a gene, for example, the number of reads that sample each position of the gene's exons, or the count of spliced reads that span a splice junction. The second strategy estimates the transcript abundance first and compares transcription profiles explicitly at the transcript level. This strategy may provide direct insights into differentiated transcripts, but the accuracy of abundance estimation is often concerned. The last

strategy leverages the computational efficiency and the biological interpretability, by analyzing the alternative splicing events. Each local event depicts how transcript sequences diverge at a locus in a gene. These event-based approaches typically have higher accuracy of transcript isoform identification and variant abundance estimation, and hence are more precise than transcript-based approaches.

Copyright © Yin Hu, 2013.

Chapter 8 Conclusion

This dissertation has presented a comprehensive framework for transcriptome analysis using high-throughput RNA sequencing data. Without dependence on any transcript annotation, the framework has the ability to discover novel exons, splice junctions and genome modifications that are not cataloged in existing database. This purely data-driven strategy also avoids the risk of missing unannotated transcripts that may code clinically important proteins or introducing unnecessary gene models that may confound abundance estimation, both may bias the result of the differential transcription analysis.

The read pairing in paired-end sequencing strategy allows higher transcriptome coverage and more validating information for resolving ambiguous read mapping. Existing methods typically use the expected adjacency between the mapped end read positions to guide the choice of best read alignments [Au et al., 2010] or use the pairing to help identify transcript structure [Trapnell et al., 2010]. The inner portion between two ends if often confounded with alternative splicing and its coverage is often not well employed. The first stage in the framework then resolve the alignment of the inner portion together with those of the sequenced end reads, as an entirety referred to as the *transcript fragment alignment*. The construction of unknown inner portions is enabled by path searching in the genome-wide splice graph reconstructed from the read alignments. All splice junctions, end read alignments and inner portion alignments are then summarized in a probabilistic model to identify the globally best

alignment maximizing the joint probability of the entire data.

The core DiffSplice algorithm consists of four major components — a probabilistic model that maps paired-end RNA-seq reads to the reference genome and infers the full transcript fragment alignments, the reconstruction of the genome-wide splice graph from the fragment alignments and the identification of alternative splicing modules, the inference of splicing isoform abundance followed by statistical differential transcription test, and a exploratory clustering scheme for consistent transcription pattern discovery. The core advance of this framework as compared to existing methodologies is the development of the divide-and-conquer approach that automatically localizes the difference between transcriptomes into Alternative Splicing Modules (ASMs) where transcript isoforms diverge.

The framework directly starts from the output sequence files out of RNA-seq experiments. Sequenced RNA-seq reads are first mapped to the reference genome using a short read aligner. For paired-end reads, MapPER is applied to find alignments for the entire transcript fragments based on the distribution of insert-size. This further consolidates the prediction of splice junctions and increases the read coverage on the transcriptome.

From the union of the RNA-seq alignments from all samples, the genome-wide expression-weighted splice graph (ESG) is constructed to summarize the expression and splicing information on the genome in the given dataset. This unified graph provides a survey of all possible alternative splicing and transcription events that may be present in any sample, condensing all samples with the same graph representation but distinguishing them by the graph weights. The ESG is then iteratively decomposed

in a top-down manner to resolve all regions where transcripts diverge and reconvene. These regions are called the ASMs. The diverging paths in each module differentiate transcripts in this locus and will be the units for analyzing differential transcription.

For every sample, isoform abundance estimation is thereafter conducted for the diverging paths in each module, based on the read distribution observed in the sample. The absolute abundance as well as the relative proportion of every alternative transcription path are estimated. This estimation procedure works outward in a bottom-up manner, from nested, smaller modules to parent, larger modules. Subsequently the estimates for the expression of each ASM are propagated to derive an estimate for the overall gene expression.

Lastly, a statistical test is performed to evaluate the magnitude and consistency of differences in transcription across sample groups, by comparing the estimated alternative path proportions in every sample. Controlled by the false discovery rate, the significance of the observed differences is assessed through a non-parametric permutation test, alleviating the risk of inappropriate assumption on the distribution of the test statistics. In this way, the detection of splicing isoforms that are differentially expressed is localized at individual ASMs. The differential expression level for each gene is similarly evaluated by testing the estimated gene expression. (Section 5)

The scope of current differential transcription analyses is limited at the group-wise analyses, which focuses on the test of differences from one sample group to another. Each sample group corresponding to a hypothetically different condition, such as “diseased” versus “normal”. This type of analysis typically assumes a population distribution that all samples in the group should follow. This assumption,

however, may not hold in clinical experiments, especially in large scale datasets like the ones in the Cancer Genome Atlas (TCGA) project. Such datasets have less clear separation of samples, such as subtypes of breast cancer, and hence larger biological variation within a sample group as well as significantly overlapped features across sample groups. The work presented in this dissertation has extended the differential transcription analysis to the scenario where the grouping information is not available or any consistent change from one group to another is subtle. A non-parametric approach has been described utilizing the hierarchical clustering combined with the Mahalanobis distance. This clustering scheme allows the discovery of prominent clusters of samples that exhibit consistent pattern of transcription on some genes, even without knowledge of the number of clusters. A statistical measurement is further developed to score and rank the clusters according to classification information they may possess.

On the basis of the transcript fragment alignments and the reconstructed genome-wide splice graph, the differential transcription analysis performs isoform abundance estimation and statistical tests at the alternative splicing level. The unique design that distinguishes this framework is the automatic identification of alternative splicing modules. Compared to the approaches based on the full transcripts, this module-based analysis may achieve higher sensitivity and specificity because it circumvents the need of full transcript reconstruction and quantification, both having been found computationally unstable and inaccurate under current technologies of short read sequencing. Because the transcript isoform identification and the quantification are local, the effects of sampling biases in the RNA-seq experiments are also alleviated,

which will also improve the precision of the differential transcription analysis. Biologically, the alternative splicing modules highlights the loci where transcripts in a gene diverge, which may help the interpretation of differential transcription by answering questions such as which sequences might be responsible for the change of phenotype.

The described framework has been evaluated extensively using a series of simulation studies and real RNA-seq datasets. Synthetic data generated by mimicking practical sequencing procedures on human transcriptome has justified the accuracy of alternative splicing variant quantification under various sampling depths and sampling biases. Comparison with other state-of-the-art methods has demonstrated the superior sensitivity and specificity of the described framework in calling genes transcribed differently in different conditions. On the lung differentiation dataset and the breast cancer dataset, the DiffSplice framework successfully identified hundreds of genes that may perform key function in cell development and cancer type distinction, including genes with novel splicing isoforms and novel insertion/deletions. The application of the framework on the 728 breast cancer samples further demonstrated the scalability of the analysis. Besides a list of clinically significant genes that consistently shift transcription from subtype to subtype, the framework also discovered a list of sample clusters that highlight the alternative splicing events which may possess high value in subtype definition and classification.

The software package of the methods described in this dissertation will be open-source, released and maintained at `http://www.netlab.uky.edu/p/bioinfo/` and freely available to the research community. The pipeline will take the RNA-seq read files in standard FASTA or FASTQ format as input and generates both table results

that record the expression profiles and transcription differences and GTF tracks that can be visualized through the genome browser. The whole software suit will be finely tuned and optimized for both computational efficiency and analyzing accuracy, serving as a highly competitive pipeline for comprehensive studies on differential transcriptome expression.

Empowered by the advancement of sequencing technologies, the role of sequencing data in any biomedical research/application is becoming more and more prominent. Three major characteristics can be envisioned for future sequencing data analysis. First, the size of data will be skyrocketing as the sequencing price keeps dropping. Tens of samples per study is turning common, and datasets with thousands of samples have become available. Scalability will be even more emphasized in the evaluation of computational approaches. Second, the complexity of data may exhibit exponential increase as sampling depth increases and knowledge expands. This complexity has also been demonstrated by the considerable discrepancy among different methods on same datasets. Methodology designs are facing unforeseen challenges in leveraging sensitivity and specificity, and in maintaining robust performance across different datasets. Third, the integration of various data sources will become more and more important. Such integrations, such as combining DNA data and RNA data, may benefit the solution of well-defined problems such as structural variation detection as well as establish connections between fundamentally related domains that have not been linked functionally such as the genome aberrations and mRNA expression.

Copyright © Yin Hu, 2013.

Bibliography

Ensembl Genome Browser. http://useast.ensembl.org/index.html. 82

NCBI Reference Sequence (RefSeq). http://www.ncbi.nlm.nih.gov/RefSeq. 82

http://www.wikipedia.org/. http://www.wikipedia.org/. ix, 15, 23

A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten. Global and unbiased detection of splice junctions from rna-seq data. *Genome Biol.*, 11:R34, 2010. 160

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 2010. 84, 127

Y. W. Asmann, A. Hossain, B. M. Necela, S. Middha, K. R. Kalari, Z. Sun, H.-S. Chai, D. W. Williamson, D. Radisky, G. P. Schroth, J.-P. A. Kocher, E. A. Perez, and E. A. Thompson. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res.*, 39(15):e100, 2011. 160

K. F. Au, H. Jiang, L. Lin, Y. Xing, and W. H. Wong. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic Acids Res.*, 38(14):4570–8, 2010. 25, 27, 47, 50, 54, 67, 157, 173

J. Bainbridge, H. Jia, A. Bagherzadeh, D. Selwood, R. Ali, and I. Zachary. A peptide encoded by exon 6 of vegf (eg3306) inhibits vegf-induced angiogenesis in vitro and

ischaemic retinal neovascularisation in vivo. *Biochemical and Biophysical Research Communications*, 302(4):793–9, 2003. 77

M. F. Berger, J. Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L. A. Johnson, J. Robinson, R. G. Verhaak, C. Sougnez, R. C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D. E. Fisher, G. Getz, M. Meyerson, D. B. Jaffe, S. B. Gabriel, E. S. Lander, R. Dummer, A. Gnirke, C. Nusbaum, and L. A. Garraway. Integrative analysis of the melanoma transcriptome. *Genome Res.*, 20:413–427, 2010. 4, 25, 159

I. Birol, S. Jackman, C. Nielsen, J. Qian, R. Varhol, G. Stazyk, R. Morin, Y. Zhao, M. Hirst, J. Schein, D. Horsman, J. Connors, R. Gascoyne, M. Marra, and S. Jones. *De novo* transcriptome assembly with abyss. *Bioinformatics*, 25:2872–2877, 2009. 66

D. L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annu. Rev. Biochem.*, 72:291336, 2003. 17

R. Bohnert and G. Rätsch. rquant.web: a tool for rna-seq-based transcript quantitation. *Nucleic Acids Res.*, 38(Suppl 2):W348–W351, 2010. 83, 104

B. Bolstad, R. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. 84

F. D. Bona, S. Ossowski, K. Schneeberger, and G. Rtsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):174–180, 2008. 157

R. Brown, L. Reinke, M. Damerow, D. Perez, L. Chodosh, J. Yang, and C. Cheng. Cd44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *The Journal of Clinical Investigation*, 121(3):1064–1074, 2011. 78

A. Buchsbaum, H. Kaplan, A. Rogers, and J. Westbrook. A new, simpler linear-time dominators algorithm. *ACM Trans. Program. Lang. Syst.*, 20(6):1265–1296, 1998. 72, 73

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(94), 2010. 84, 135

J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. Allpaths: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18(5):810–20, 2008. 163

D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. *The 16th ACM SIGKDD conference on Knowledge Discovery and Data mining*, 2010. 84

D. Clark. *Molecular biology.* Amsterdam: Elsevier Academic Press, 2005. 17

T. Clark, C. Sugnet, and M. Ares. Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296:907–910, 2002. 20, 56

A. Cox. Novoalign. ELAND: Efficient large-scale alignment of nucleotide databases. Illumina, San Diego., 2007. 156

C. Dai, W. Li, J. Liu, and X. J. Zhou. Integrating many co-splicing networks to reconstruct splicing regulatory modules. *BMC Systems Biology*, 6(Suppl 1):S17, 2012. 139

A. P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39 (1):1–38, 1977. 36

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Lalo, C. L. Gall, B. Schaffer, S. L. Crom, M. Guedj, and F. J. on behalf of The French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Brief Bioinform*, 2012. doi: 10.1093/bib/bbs046. 84

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. 157

N. A. Faustino and T. A. Cooper. Pre-mrna splicing and human disease. *Genes & Dev.*, 17:419437, 2003. 19

J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. *J Comput Biol.*, 18(3):305–21, 2011. 167

J. Ferrante, K. Ottenstein, and J. Warren. The program dependence graph and its use in optimization. *ACM Trans. Program. Lang. Syst.*, 9(3):319–349, 1987. 72

R.-H. Fu, S.-P. Liu, S.-J. Huang, H.-J. Chen, P.-R. Chen, Ya-Hsien, Y.-C. Ho, W.-L. Chang, C.-H. Tsai, W.-C. Shyu, and S.-Z. Lin. Aberrant alternative splicing events in parkinson's disease. *Cell Transplantation*, 22(4):653661, 2013. 20

M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. *Nat. Methods*, 8: 469–477, 2011. 156, 169

H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck. Fusionmap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, 27(14):1922–1928, 2011. 160

M. Grabherr, B. Haas, M. Yassour, J. Levin, D. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–52, 2011. 66, 164

M. Griffith, O. Griffith, J. Mwenifumbo, R. Goya, A. Morrissy, R. D. Morin, R. Corbett, M. Tang, Y.-C. Hou, T. Pugh, G. Robertson, S. Chittaranjan, A. Ally, J. Asano, S. Chan, H. Li, H. McDonald, K. Teague, Y. Zhao, T. Zeng, A. Delaney, M. Hirst, G. Morin, S. Jones, I. Tai, and M. Marra. Alternative expression analysis by rna sequencing. *Nat. Methods*, 7:843–847, 2010. 59, 63

M. Guttman, M. Garber, J. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. Koziol, A. Gnirke, C. Nusbaum, J. Rinn, E. Lander, and A. Regev. Ab initio

reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat. Biotechnol.*, 28:503–510, 2010. 167, 169

S. Haile and M. D. Sadar. Androgen receptor and its splice variants in prostate cancer. *Cellular and Molecular Life Sciences*, 68(24):39713981, 2011. 20

Q. He and D.-Y. Lin. A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8, 2010. 84

S. Heber, M. Alekseyev, S. Sze, H. Tang, and P. Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, 18 (suppl 1):S181–S188, 2002. 66, 167

C. Hercus. Novoalign. www.novocraft.com. 156

D. Hiller, H. Jiang, W. Xu, and W. H. Wong. Identifiability of isoform deconvolution from junction arrays and rna-seq. *Bioinformatics*, 25:3056–3059, 2009. 83

F. Hormozdiari, C. Alkan, E. E. Eichler, and S. Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, 19:1270–1278, 2009. 159

Y. Hu, K. Wang, X. He, D. Chiang, J. Prins, and J. Liu. A probabilistic framework for aligning paired-end rna-seq data. *Bioinformatics*, 26:1950–1957, 2010. 66, 67

Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, A. Monroy, P.-F. Kuan, S. M. Hammond, L. Makowski, S. H. Randell, D. Y. Chiang, D. N. Hayes, C. Jones, Y. Liu, J. F. Prins, and J. Liu. Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic Acids Res.*, 2012. 63, 167

Y. Huang, Y. Hu, C. D. Jones, J. N. MacLeod, D. Y. Chiang, Y. Liu, J. F. Prins, and J. Liu. A robust method for transcript quantification with rna-seq data. *16th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2012. 57, 83, 84

M. K. Iyer, A. M. Chinnaiyan, and C. A. Maher. Chimerascan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011. 160

H. Jiang and W. Wong. Statistical inference for isoform expression in rna-seq. *Bioinformatics*, 25:1026–1032, 2009a. 80, 86, 87

H. Jiang and W. H. Wong. Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, 25:1026–1032, 2009b. doi: 10.1093/bioinformatics/btp113. 83

Y. Katz, E. Wang, E. Airoldi, and C. Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–1015, 2010. 59, 63

W. J. Kent. Blatthe blast-like alignment tool. *Genome Res.*, 12:656664, 2002. 131

D. Kim and S. L. Salzberg. Tophat-fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, 12:R72, 2011. 160

M. Kinsella, O. Harismendy, M. Nakano, K. A. Frazer, and V. Bafna. Sensitive gene fusion detection using ambiguously mapping rna-seq read pairs. *Bioinformatics*, 27(8):1068–1075, 2011. 160

L. Klebanov and A. Yakovlev. How high is the level of technical noise in microarray data? *Biol Direct.*, 2:9, 2007. 22

G. Koscielny, V. L. Texier, C. Gopalakrishnan, V. Kumanduri, J.-J. Riethoven, F. Nardone, E. Stanley, C. Fallsehr, O. Hofmann, M. Kull, E. Harrington, S. Boue, E. Eyras, M. Plass, F. Lopez, W. Ritchie, V. Moucadel, T. Ara, H. Pospisil, A. Herrmann, J. G. Reich, R. Guig, P. Bork, M. von Knebel Doeberitz, J. Vilo, W. Hide, R. Apweiler, T. A. Thanaraj, and D. Gautheret. Astd: The alternative splicing and transcript diversity database. *Genomics.*, 93(3):213–20, 2009. viii, 49

M. Krzywinski, J. Schein, nan Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Res*, 19:1639–1645, 2009. xi, 53

S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 1951. doi:10.1214/aoms/1177729694. 95

T. Kwan, D. Benovoy, C. Dias, S. Gurd, C. Provencher, P. Beaulieu, T. Hudson, R. Sladek, and J. Majewski. Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.*, 40:225–31, 2008. 17, 56

V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron. Exact transcriptome reconstruction from short sequence reads. *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, 2008. 83

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9:357–359, 2012. 156

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10:R25, 2009. 156

S. Lee, E. Cheran, and M. Brudno. A robust framework for detecting structural variations in a genome. *Bioinformatics*, 24(13):i59–i67, 2008. 159

B. Li and C. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinf.*, 12:323, 2011. 83

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, 2010a. doi: 10.1093/bioinformatics/btp692. 83, 127

H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–60, 2009. 156

H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, 2008a. 156

H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1821–1858, 2008b. 34

J. Li, H. Jiang, and W. H. Wong. Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biol.*, 11, 2010b. 84

R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, 20:265–272, 2010c. 66

W. Li, J. Feng, and T. Jiang. Isolasso: A lasso regression approach to rna-seq based transcriptome assembly. *J Comput Biol.*, 18(11):1693–1707, 2011a. 83, 84, 167, 169

Y. Li, J. Chien, D. I. Smith, and J. Ma. Fusionhunter: identifying fusion transcripts in cancer using paired-end rna-seq. *Bioinformatics*, 27(12):1708–1710, 2011b. 160

J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991. doi:10.1109/18.61115. 95

A. J. Lopez. Alternative splicing of pre-mrna: Developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, 32:279305, 1998. 17

G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res.*, 21:936–939, 2011. 156

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Trans. Comput. Biol. Bioinform.*, 1(1):24–45, 2004. 138

C. A. Maher, N. Palanisamy, J. C. Brenner, X. Cao, S. Kalyana-Sundaram, S. Luo, I. Khrebtukova, T. R. Barrette, C. Grasso, J. Yu, R. J. Lonigro, G. Schroth, C. Kumar-Sinha, and A. M. Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA*, 106(30):12353–12358, 2009. viii, xi, 4, 25, 28, 30, 52, 53, 54, 55, 159

S. Marguerat, B. T. Wilhelm, and J. Bahler. Next-generation sequencing: applications beyond genomes. *Biochemical Society Transactions*, 36(5):1091–1096, 2008. 152

J. Martin, V. M. Bruno, Z. Fang, X. Meng, M. Blow, T. Zhang, G. Sherlock, M. Snyder, and Z. Wang. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded rna-seq reads. *BMC Genomics*, 11:663, 2010. 164

J. A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12:671–682, 2011. 164

A. J. Matlin, F. Clark, and C. W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell. Biol.*, 6:386398, 2005. 17

National Cancer Institute. *Targeted Cancer Therapies*. `http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted`. 4

National Center for Biotechnology Information. *National Center for Biotechnology Information*. `http://www.ncbi.nlm.nih.gov`. 21

National Human Genome Research Institute. *National Human Genome Research Institute*. `http://www.genome.gov`. 21

A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. F. Sun, M. Griffith, A. H. Moussavi, J. Senz, N. Melnyk, M. Pacheco, M. A. Marra, M. Hirst, T. O. Nielsen, S. C. Sahinalp, D. Huntsman, and S. P. Shah. defuse: An algorithm for gene fusion discovery in tumor rna-seq data. *PLoS Comput. Biol.*, 7(5):e1001138, 2011a. 160

A. McPherson, C. Wu, I. Hajirasouliha, F. Hormozdiari, F. Hach, A. Lapuk, S. Volik, S. Shah, C. Collins, and S. C. Sahinalp. Comrad: detection of expressed rearrangements by integrated analysis of rna-seq and low coverage genome sequence data. *Bioinformatics*, 27(11):1481–1488, 2011b. 160

A. McPherson, C. Wu, A. W. Wyatt, S. Shah, C. Collins, and S. C. Sahinalp. nfuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.*, 22:2250–2261, 2012. 160

P. Medvedev and M. Brudno. Ab initio whole genome shotgun assembly with mated short reads. In *RECOMB*, pages 50–64, 2008. 159

P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, 6:S13–S20, 2009. 159

R. Muraoka-Cook, M. Sandahl, K. Strunk, L. Miraglia, C. Husted, D. Hunter, K. Elenius, L. Chodosh, and H. S. Earp. Erbb4 splice variants cyt1 and cyt2 differ by 16 amino acids and exert opposing effects on the mammary epithelium in vivo. *Molecular and Cellular Biology*, 29(18):4935–48, 2009. 77

M. Nicolae, S. Mangul, I. I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011. 83

M. Okoniewski and C. Miller. Comprehensive analysis of affymetrix exon arrays using bioconductor. *PLoS Comput. Biol.*, 4(2):e6, 2008. 21, 56

M. Olejniczak, P. Galka, and W. Krzyzosiak. Sequence-non-specific effects of rna interference triggers and microrna regulators. *Nucleic Acids Res.*, 38(1):1–16, 2010. 104

Q. Pan, O. Shai, L. Lee, B. Frey, and B. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40:1413–1415, 2008. 17, 22, 56

K. Pingali and G. Bilardi. Optimal control dependence computation and the roman chariots problem. *ACM Trans. Program. Lang. Syst.*, 19(3):462–491, 1997. 72

R. Pio and L. M. Montuenga. Alternative splicing in lung cancer. *Journal of Thoracic Oncology*, 4(6):674678, 2009. 20

M. G. Poulos, R. Batra, K. Charizanis, and M. S. Swanson. Developments in rna splicing and disease. *Cold Spring Harbor Perspectives in Biology*, 3(1), 2011. 20

H. Richard, M. H. Schulz, M. Sultan, A. Nrnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. A. Haas, and M.-L. Yaspo. Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Res.*, 38:e112, 2010. doi: 10.1093/nar/gkq041. 83

A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biol.*, 12(3):R22, 2011. 83, 84, 104

G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard,

Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of rna-seq data. *Nat. Methods*, 7:909–912, 2010. 160, 164

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010. 84

S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. Shrimp: Accurate mapping of short color-space reads. *PLoS Comput. Biol.*, 5(5): e1000386, 2009. 156

G. Russo, C. Zegar, and A. Giordano. Advantages and limitations of microarray technology in human cancer. *Oncogene*, 22:64976507, 2003. 20

M. Sammeth. Complete alternative splicing events are bubbles in splicing graphs. *J Comput Biol.*, 16(8):1117–1140, 2009. 167

A. Sboner, L. Habegger, D. Pflueger, S. Terry, D. Z. Chen, J. S. Rozowsky, A. K. Tewari, N. Kitabayashi, B. J. Moss, M. S. Chee, F. Demichelis, M. A. Rubin, and M. B. Gerstein. Fusionseq: a modular framework for finding gene fusions by analyzing paired-end rna-sequencing data. *Genome Biol.*, 11:R104, 2010. 160

M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012. 164

S. Shen, J. Park, J. Huang, K. A. Dittmar, Z. 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 Res.*, 2012. 59, 63

T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA*, 100(26):15776–15781, 2003. 21

J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and nan Birol. Abyss: A parallel assembler for short read sequence data. *Genome Res.*, 19:1117–1123, 2009. 163

D. Singh, C. Orellana, Y. Hu, C. Jones, Y. Liu, D. Chiang, J. Liu, and J. Prins. Fdm: A graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics*, 27:2633–2640, 2011. 59, 63, 66, 99, 106, 119, 167

S. Srivastava and L. Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Res.*, pages 1–15, 2010a. doi: 10.1093/nar/gkq670. 80, 104

S. Srivastava and L. Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Res.*, 38(17):e170, 2010b. doi: 10.1093/nar/gkq041. 83

O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential rna-transcript expression from read counts. *Nature Precedings*, 2010. 59, 63

M. Sultan, M. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321: 956–960, 2008. 17, 56

Y. Surget-Groba and J. I. Montoya-Burgos. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.*, 20(10):1432–40, 2010. 164

C. Tammaro, M. Raponi, D. I. Wilson, and D. Baralle. Brca1 exon 11 alternative splicing, multiple functions and the association with cancer. *Biochemical Society Transactions*, 40(4):768772, 2012. 20

J. Tazia, N. Bakkoura, and S. Stamm. Alternative splicing and disease. *Biochim Biophys Acta.*, 1792(1):1426, 2009. 19

C. Trapnell, L. Pachter, and S. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25:1105–1111, 2009a. 67

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009b. 25, 27, 157

C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28:511–515, 2010. doi: 10.1038/nbt.1621. 19, 56, 57, 63, 66, 83, 84, 127, 167, 169, 173

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl. Acad. Sci. USA*, 99(22):1403114036, 2002. 22

E. Turro, S.-Y. Su, A. Goncalves, L. J. Coin, S. Richardson, and A. Lewin. Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biol.*, 12:R13, 2011. 84

V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98(9):5116–21, 2001. 65, 96, 98, 111

V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995. 21

E. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008. 17, 19, 56

G. Wang and T. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, 8:749–761, 2007. 20, 56

K. Wang, D. Singh, Z. Zeng, Y. Huang, S. Coleman, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu. Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38 (18):178, 2010a. 29, 66, 111, 157, 160

Y. Wang, D. W. Chan, V. W. Liu, P. Chiu, and H. Y. Ngan. Differential functions of growth factor receptor-bound protein 7 (grb7) and its variant grb7v in ovarian carcinogenesis. *Clinical Cancer Research*, 16(9):25292539, 2010b. 20

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63, 2009. 22, 57, 152, 164

Wu and C. F. Jeff. On the convergence properties of the em algorithm. *Annals of Statistics*, 11(1):95–103, 1983. 36

J. Wu, M. Akerman, S. Sun, W. McCombie, A. Krainer, and M. Zhang. Splicetrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, pages btr508v1–btr508, 2011a. 59, 63

T. D. Wu and S. Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010. 157

Z. Wu, X. Wang, and X. Zhang. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics*, 27:502–508, 2011b. 83

L. Xi, A. Feber, V. Gupta, M. Wu, A. Bergemann, R. Landreneau, V. Litle, A. Pennathur, J. Luketich, and T. Godfrey. Whole genome exon arrays identify differen-

tial expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res.*, 36(20):6535–6547, 2008. 21, 57

X. Xu, D. Yang, J.-H. Ding, W. Wang, P.-H. Chu, N. D. Dalton, H.-Y. Wang, J. R. Bermingham, Z. Ye, F. Liu, M. G. Rosenfeld, J. L. Manley, J. Ross, J. Chen, R.-P. Xiao, H. Cheng, and X.-D. Fu. Asf/sf2-regulated camkii alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell*, 120(1):5972, 2005. 20

K. Yap and E. V. Makeyev. Regulation of gene expression in mammalian nervous system through alternative pre-mrna splicing coupled with rna quality control mechanisms. *Molecular and Cellular Neuroscience*, 2013. 20

D. Yorukoglu, F. Hach, L. Swanson, C. C. Collins, I. Birol, and S. C. Sahinalp. Dissect: detection and characterization of novel structural alterations in transcribed sequences. *Bioinformatics*, 28(12):i179–87, 2012. 160

M. Zaharia, W. J. Bolosky, K. Curtis, A. Fox, D. Patterson, S. Shenker, I. Stoica, R. M. Karp, and T. Sittler. Faster and more accurate sequence alignment with snap. arXiv:1111.5572v1, 2011. 156

D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18(5):821–829, 2008. 163

D. R. Zerbino, G. K. McEwen, E. H. Margulies, and E. Birney. Pebble and rock band: Heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*, 4(12):e8407, 2009. 163

J. Zhang, K. Lu, Y. Xiang, M. Islam, S. Kotian, Z. Kais, C. Lee, M. Arora, H. wen Liu, J. D. Parvin, and K. Huang. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput. Biol.*, 8(8): e1002656, 2012. 138

Vita

Name

Yin Hu

Education

M.Sc., Statistics	May, 2013
University of Kentucky	Lexington, KY, USA
B.Eng., Computer Science	July, 2008
University of Science and Technology of China	Hefei, China

Publications

1. Yan Huang, **Yin Hu**, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang, Yufeng Liu, Jan F. Prins and Jinze Liu. A Robust Method for Transcript Quantification with RNA-seq Data. ***Journal of Computational Biology***, March 2013, 20(3): 167-187.

2. **Yin Hu**, Yan Huang, Ying Du, Christian F. Orellana, Darshan Singh, Amy Johnson, Anais Monroy, Pei-Fen Kuan, Scott Hammond, Liza Makowski, Scott Randell, Derek Y. Chiang, David Hayes, Corbin D. Jones, Yufeng Liu, Jan F. Prins and Jinze Liu. DiffSplice: the Genome-Wide Detection of Differential Splicing Events with RNA-seq. ***Nucleic Acids Research***, 2012, doi: 10.1093/nar/gks1026.

3. Yan Huang, **Yin Hu**, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang,

Yufeng Liu, Jan F. Prins and Jinze Liu. A Robust Method for Transcript Quantification with RNA-seq Data. ***16th Annual International Conference on Research in Computational Molecular Biology (RECOMB)***, Apr 21 - Apr 24, 2012, Barcelona, Spain.

4. Darshan Singh, Christian F. Orellana, **Yin Hu**, Corbin D. Jones, Yufeng Liu, Derek Y. Chiang, Jinze Liu and Jan F. Prins. FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. ***Bioinformatics***, 2011, 27(19): 2633-640, doi: 10.1093/bioinformatics/btr45.

5. **Yin Hu**, Kai Wang, Xiaping He, Derek Y. Chiang, Jan F. Prins and Jinze Liu. A Probabilistic Framework for Aligning Paired-end RNA-seq Data. ***Bioinformatics***, 2010, 26(16):1950-1957, doi:10.1093/bioinformatics/btq336.

6. Jizhou Gao, **Yin Hu**, Jinze Liu and Ruigang Yang. Unsupervised Learning of High-order Structural Semantics from Images. ***IEEE 12th International Conference on Computer Vision (ICCV)***, Sept 27 - Oct 4, 2009, Kyoto, Japan, pp.2122-2129, doi:10.1109/ICCV.2009.5459465.

Conference Presentations

1. **Yin Hu**, Jan F. Prins and Jinze Liu. DiffSplice: the Genome-Wide Detection of Differential Splicing Events with RNA-seq. ***ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB)***, Washington DC, USA, September, 2013. (Highlights talks)

2. **Yin Hu**, Yan Huang, Ying Du, Christian F. Orellana, Darshan Singh, Amy Johnson, Anais Monroy, Pei-Fen Kuan, Scott Hammond, Liza Makowski, Scott Randell, Derek Y. Chiang, David Hayes, Corbin D. Jones, Yufeng Liu, Jan F. Prins and Jinze Liu. DiffSplice: the Genome-Wide Detection of Differential Splicing Events with RNA-seq. ***20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)***, Long Beach, USA, July, 2012. (Recommended by Faculty of 1000)

3. Yan Huang, **Yin Hu**, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang, Yufeng Liu, Jan F. Prins and Jinze Liu. A Robust Linear Framework for Transcript Quantification using MultiSplice Features. ***20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)***, Long Beach, USA, July, 2012. (Oral presentation)

4. **Yin Hu**, Yan Huang, Derek Y. Chiang, Corbin D. Jones, Jan F. Prins and Jinze Liu. An *Ab Initio* Method for Differential Transcriptome Analysis. ***UT-ORNL-KBRIN Bioinformatics Summit 2012***, Louisville, USA, March, 2012.

5. Yan Huang, **Yin Hu**, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang, Yufeng Liu, Jan F. Prins and Jinze Liu. A Linear Framework for Transcript Quantification from RNA-seq Data. ***UT-ORNL-KBRIN Bioinformatics Summit 2012***, Louisville, USA, March, 2012. (Oral presentation)

6. **Yin Hu**, Yan Huang, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang,

D. Neil Hayes, Jan F. Prins and Jinze Liu. Detection and Quantification of Differentially Expressed Genes using RNA-seq. ***The 2011 Southeast Regional IDeA Meeting***, New Orleans, USA, September, 2011. (Oral presentation)

7. Yan Huang, **Yin Hu**, Matthew S. Hestand, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang, Yufeng Liu, Jan F. Prins and Jinze Liu. A Robust Method for Transcript Quantification with RNA-seq Data. ***ISMB Special Interest Group on High Throughput Sequencing Analysis and Algorithms (HiTSeq)***, Vienna, Austria, July, 2011.

8. Darshan Singh, Christian F. Orellana, **Yin Hu**, Corbin D. Jones, Yufeng Liu, Derek Y. Chiang, Jinze Liu and Jan F. Prins. FDM: A Graph-based Statistical Method to Detect Differential Transcription from RNA-seq Data. ***ISMB Special Interest Group on High Throughput Sequencing Analysis and Algorithms (HiTSeq)***, Vienna, Austria, July, 2011. (Oral presentation)

9. **Yin Hu**, Kai Wang, Xiaping He, Derek Y. Chiang, Jan F. Prins and Jinze Liu. A Probabilistic Framework for Accurate Detection of Gene Fusion Events with Paired-End RNA-seq Reads. ***ISMB Special Interest Group on High Throughput Sequencing Analysis and Algorithms (HiTSeq)***, Boston, USA, July, 2010. (Oral presentation)

10. **Yin Hu** and Jinze Liu. Detecting Gene Fusion with Paired-End Reads. ***The Annual Conference of Midsouth Computational Biology*** & ***Bioinformatics Society (MCBIOS)***, Jonesboro, USA, February, 2010. (Oral

presentation)

Software

DiffSplice: detection and quantification of differential transcriptome expression using RNA-seq data.

- http://www.netlab.uky.edu/p/bioinfo/DiffSplice

MapPER: probabilistic alignment and full fragment reconstruction of paired-end RNA-seq data.

- http://www.netlab.uky.edu/p/bioinfo/MapPER

Awards

Kentucky Opportunity Fellowship	2011
Huawei Scholarship	2007
Autonomous Enrollment of USTC	2004