2015

NOVEL COMPUTATIONAL METHODS FOR TRANSCRIPT RECONSTRUCTION AND QUANTIFICATION USING RNA-SEQ DATA

Yan Huang
University of Kentucky, yhu233@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 needed written permission statement(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) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free 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 an 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 thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Yan Huang, Student

Dr. Jinze Liu, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

NOVEL COMPUTATIONAL METHODS FOR TRANSCRIPT RECONSTRUCTION AND QUANTIFICATION USING 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
Yan Huang
Lexington, Kentucky

Director: Dr. Jinze Liu, Associate Professor of Computer Science
Lexington, Kentucky 2015

ABSTRACT OF DISSERTATION

NOVEL COMPUTATIONAL METHODS FOR TRANSCRIPT RECONSTRUCTION AND QUANTIFICATION USING RNA-SEQ DATA

The advent of RNA-seq technologies provides an unprecedented opportunity to precisely profile the mRNA transcriptome of a specific cell population. It helps reveal the characteristics of the cell under the particular condition such as a disease. It is now possible to discover mRNA transcripts not cataloged in existing database, in addition to assessing the identities and quantities of the known transcripts in a given sample or cell. However, the sequence reads obtained from an RNA-seq experiment is only a short fragment of the original transcript. How to recapitulate the mRNA transcriptome from short RNA-seq reads remains a challenging problem. We have proposed two methods directly addressing this challenge. First, we developed a novel method MultiSplice to accurately estimate the abundance of the well-annotated transcripts. Driven by the desire of detecting novel isoforms, a max-flow-min-cost algorithm named Astroid is designed for simultaneously discovering the presence and quantities of all possible transcripts in the transcriptome. We further extend an *ab initio* pipeline of transcriptome analysis to large-scale dataset which may contain hundreds of samples. The effectiveness of proposed methods has been supported by a series of simulation studies, and their application on real datasets suggesting a promising opportunity in reconstructing mRNA transcriptome which is critical for revealing variations among cells (e.g. disease vs. normal).

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

Author's signature: Yan Huang

Date: January 23, 2015

NOVEL COMPUTATIONAL METHODS FOR TRANSCRIPT RECONSTRUCTION AND QUANTIFICATION USING RNA-SEQ DATA

By
Yan Huang

Director of Dissertation: Jinze Liu

Director of Graduate Studies: Miroslaw Truszczyński

Date: January 23, 2015

ACKNOWLEDGMENTS

First and foremost I wish to thank my advisor, Prof. Jinze Liu, who has guided me throughout my Ph.D. study with enthusiasm, patience, caring, and immense knowledge. Besides her continuous encouragement and support, she also offered tremendous opportunities for me to attend conferences, deliver presentations. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would also like to thank my committee members: Prof. James N. MacLeod, Prof. Ruigang Yang and Prof. Nathan Jacobs for their insightful comments and suggestions. Special thanks to Prof. Jan F. Prins who is a collaborator and also the advisor during my visit in the computer science department in UNC-Chapel Hill. He always provided me with good arguments on algorithms and performance of our methods. I appreciate the experience working with him.

Last but not least, I thank my parents for their unconditional love and support, and my husband for his understanding and great help all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Uncovering the mystery of the functioning and heredity of living organisms has been and remains to be one central mission of the life sciences. *Human genetics*, the study of genes, heredity, and variation of human species, is of great interest. Genes are basic functional units that determine an individuals' unique traits. They not only decide our hair color but also holds the information of the genetic traits passing to offspring. More specifically, the research of human genetics will help unveil all fascinating mechanisms about human: *i.e.* how genetic inheritance or various characteristics takes place from parents to kids, and therefore, is highly important.

One of the major milestones of the human genetics study was the effort of Human Genome Project (HGP) which aims at identifying approximate all the $30,000$ genes in human DNA. Upon its completion in 2003, over three billion human DNA bases have been catalogued in the database, shedding lights on the study of human genetics.

Moreover, the study on human genome also led to a new era of "genomic medicine" where genetics is playing a more and more important role in the diagnosis, prognosis, and treatment of diseases that are caused by genetic abnormalities and mutations. Before genomic medicine, most diseases were defined by clinical symptoms and treated with one-fits-all treatments. This approach failed to account for individual biological background. However, nowadays more and more diseases are being defined at the molecular level, facilitating one's unique genetic information being used to improve health outcome. With the revolution of genome analysis, we are able to detect

biomarkers distinguishing between normal and diseased cells. For example the genetic characteristics responsible for tumor progression in breast cancer (e.g. $HER2+$ marker) could serve as potential drug responses and are very informative for precise and personalized treatment. Furthermore, genetic factors can also help to assess the risk of a particular disease in an individual. This is known as genetic tests. Once a high risk is confirmed, continuous monitoring and preventive measures can be taken to reduce the risk of that disease.

The great success of genetics in medical practice have stimulated the development of sequencing technology which aims to determine the sequence of entire genome, individual genes or other important molecules in a living organism. In a typical sequencing experiment, the bases of a small fragment of DNA are sequentially detected and millions of such fragments are generated from target DNA. However, only knowing the sequence is not enough, the real challenge lies in how to elucidate the connection between sequence and the gene functions, and further find out the way genes are related to phenotypes and diseases. In the past decade, many contributions have been made to answer this question, but efficient and effective computational methods are still in emerging needs. Therefore, this dissertation is dedicated to develop computational models to bridge the gap between raw sequencing data and biological findings. In this chapter, we will briefly review the biological backgrounds, introduce the problems we wish to address, and present an overview of our efforts.

1.1 Biological background

Francis Crick, the Nobel Prize winner in Physiology or Medicine in 1962, has once given a straightforward explanation of the flow of genetic information in a living organism: "DNA makes RNA and RNA makes protein" (Figure 1.1). This statement was then augmented into the central dogma of molecular biology as the following "DNA makes RNA, RNA makes proteins, proteins make us". Though a simplification, this general rule sketches the connections of all important molecules and emphasizes the order of the events in our body.

In this section, we will follow the central dogma, review the basic biological concepts and then introduce the related sequencing technology platforms for analysis.

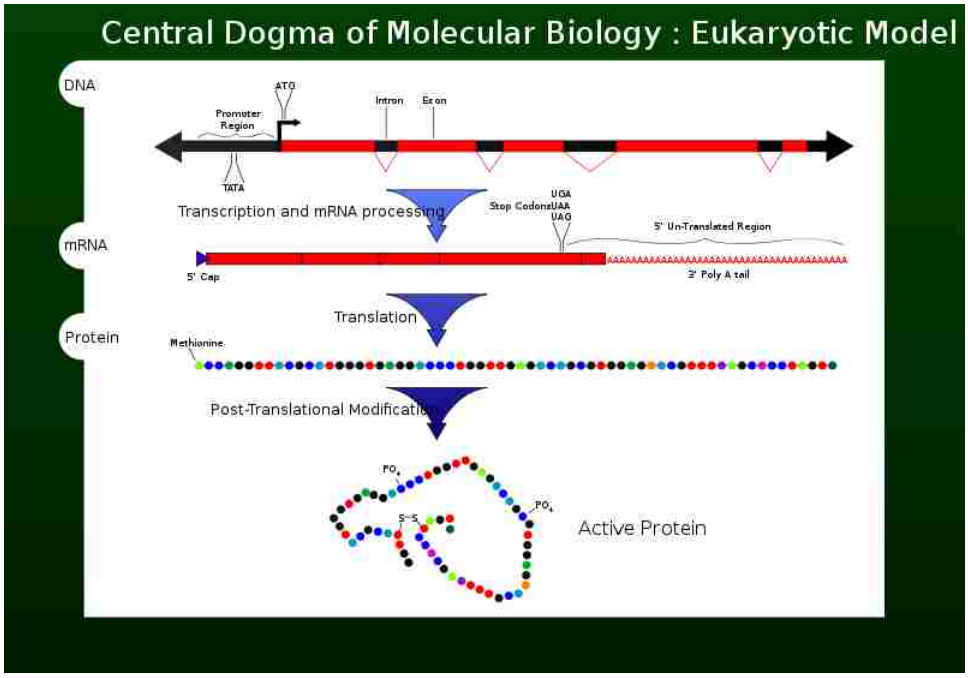
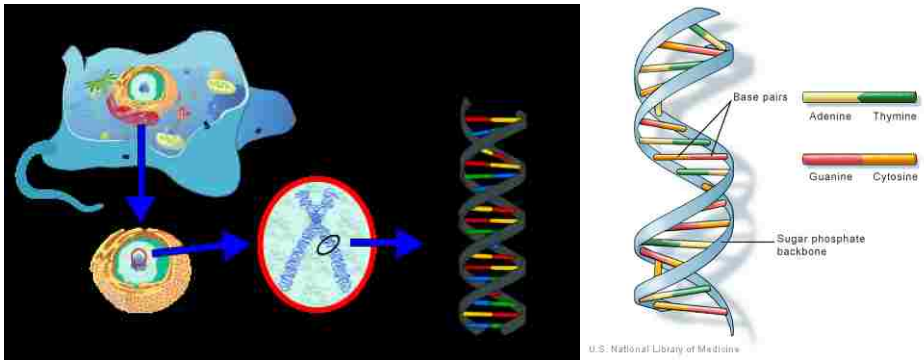


Figure 1.1: A figure illustrates the central dogma of molecular biology. Figure accommodated from Wikepedia(`www.wikepedia.com`)

.

1.1.1 The genome in living organisms

For all living organisms, cells serve as their basic building blocks. Take human body for example, it is made of trillions of cells. They support structure for the body and carry out specialized functions. Most importantly, they hold the hereditary material instructing the development and functioning of the organism (Figure 1.2a). This important information is stored in DNA (deoxyribonucleic acid). DNA is a molecule with double stranded structure as shown in Figure 1.2b. It is composed of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order, or sequence of these bases then determines the hereditary information. Human DNA contains more than 3 billion bases.



(a) (b)

Figure 1.2: (a) In cells, nuclear DNA resides within the chromosomes. (b) DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone. Figure accommodated from Wikepedia(`www.wikepedia.com`) and Genetics Home Reference(`www.ghr.nlm.nih.gov`)

The *genome* refers to the entire set of unique DNA that makes up a particular organism. On the genome, there are segments of sequences which serve as the basic physical and functional unit of heredity. They are called *genes*. In humans, genes

vary in size from a few hundred DNA bases to more than 2 million bases, and totally less than 30,000 genes are identified. For all human beings, they share more than 99% similarity at the gene-level. But the small differences (less than 1% of genes) contribute greatly to the diversity of people.

1.1.2 The mRNA transcripts and proteins

Proteins are very important molecules consisting of one or more chains amino acids. Basically, every function in a living cell depends on proteins, from antibody and enzyme to structural component and transport/storage [Phizicky et al., 2003]. Since proteins play very critical roles in a living organism, it is a long-term interest for scientists pursuing the mechanism of protein activities and deciphering their relationship with the genes. The journey from gene to protein is complex. It consists of two steps: transcription and translation (Figure 1.3).

Within a gene, the sequences that will code the final protein sequence are specified in the unit of exons (functional parts), and the rest sequences are called the introns (non-functional parts). In the process of *transcription*, introns are removed and the exons are concatenated in an mRNA transcript, following the transcription order of the gene. Each mRNA transcript then serves as a template for producing a protein. On the mRNA transcript, every contiguous three bases, called a codon, code for one particular amino acid. In the *translation*, the amino acids are assembled sequentially from the start codon to the end codon and form a protein.

However, the process of transcription can be further complicated by the mechanism of alternative splicing, through which different subsets of exons in a gene may be

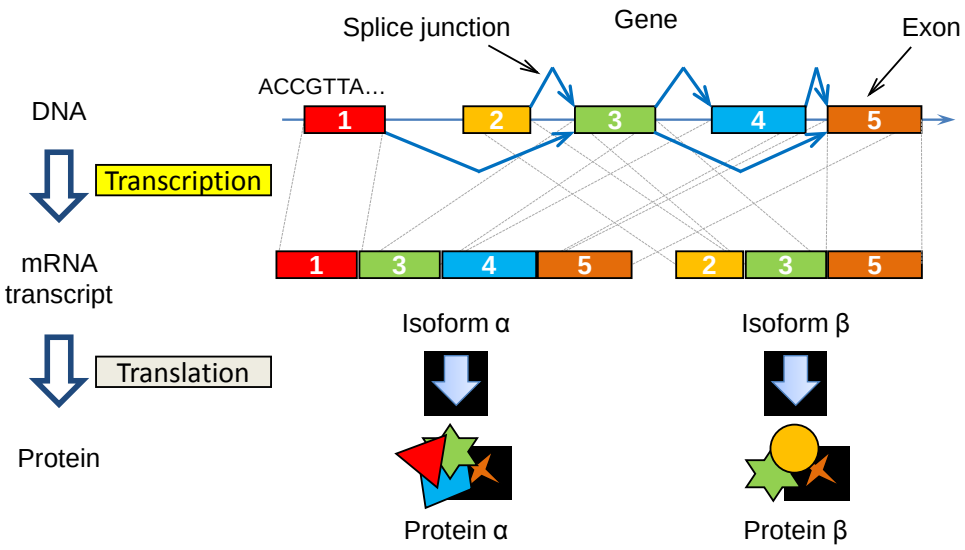


Figure 1.3: Genes and mRNA transcripts. Through alternative splicing, an important regulation process, one gene may code for multiple mRNA transcripts. Here a gene can be considered as a directed acyclic graph with vertices representing exons and directed edges representing splice junctions (introns). Different sets of exons may be retained to form different mRNA transcripts, for example, exons 1, 3, 4 and 5 form isoform α and the other set of exons form isoform β. 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.

concatenated to form different transcript isoforms [Sultan et al., 2008, Wang et al., 2008a, Pan et al., 2008a, Kwan et al., 2008]. Unlike the genome which is mostly fixed except for small changes such as mutations, the transcripts present and their individual abundance may vary in response to time and environmental factors hence may characterize the cell at a specific condition. See Figure 1.3. The analysis of the *mRNA transcriptome*, which consists of mRNA transcripts that are transcribed from the protein-coding genes then becomes a key to revealing the linkage from genotype to phenotype [Adams, 2008, Wang et al., 2008a].

By studying the transcriptome, we could find out which transcript isoforms are turned on and off in various cells or tissues (qualitative analysis). Also, the quanti-

ties of the expressed transcript isoforms could be used for analyzing their behaviors (quantitative analysis). Both studies are fundamental for downstream differential expression and transcription analysis between normal and diseased cells which help identify the biomarkers for potential drug target [Wang and Cooper, 2007]. Similarly, they could provide insight into the changes of the transcriptome at various stages of development [Wang et al., 2008a, Trapnell et al., 2010a].

1.1.3 Traditional approaches for transcriptome study

The traditional technique for studying the transcriptome is DNA Microarray technology [Clark et al., 2002, Russo et al., 2003]. In a Microarray experiment, thousands of spotted samples known as probes with known identity (pre-knowledge of sequences) are immobilized on a solid support. The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the sequences thus allowing parallel analysis for gene expression. Figure 1.4 illustrates the detailed procedure of the experiment. The sequences of interest is first purified, then PCR is used for to amplify the sequences. The core principle behind microarrays is hybridization between two DNA strands, the specific pairing of complementary nucleic acid sequences. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. A nucleic acid target is fluorescently labeled, hybridized to the sequences, and washed after hybridization. The abundance of the provided sequence is the strength of the signal which depends on the amount of targets binding to the sequence [Wiki].

The Microarray has been used as a powerful tool to measure the expression levels

of large number of genes simultaneously, which has empowered the full understanding of human genome and transcriptome.

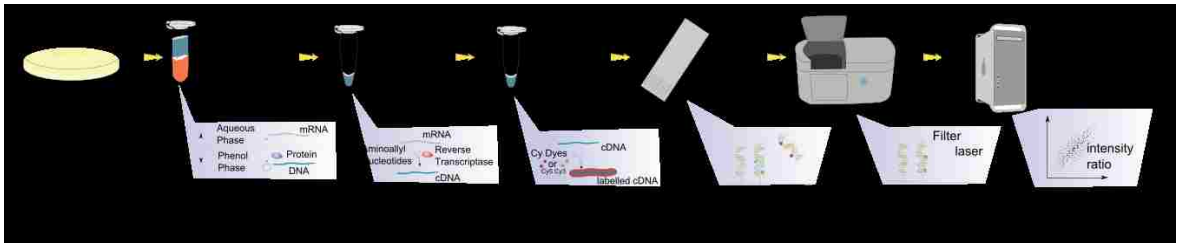


Figure 1.4: Microarray experiment steps. Figure accommodated from Wikepedia(`www.wikepedia.com`)

Besides Microarray, there exist other 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. In the experiments, the small fragments from the very beginnings (5' ends of capped transcripts) or the end (3' ends of capped transcripts) of mRNAs are extracted, reverse-transcribed to DNA, PCR amplified and sequenced. The expression level of the transcripts can be estimated by the observed counts of the sequenced fragments.

All these technologies suffer from some limitations. For example, they all require the pre-knowledge of gene/transcript sequences, which prohibits the detection of novel ones. Moreover, the dynamic range of mRNA expression levels in a cell is huge: some have only few copies while the most abundant ones may have over 10,000 copies. However, Microarray usually suffers from loss of signal at very abundant mRNAs, making it have less power on accurately quantifying these mRNAs [Wang et al., 2009b]. Microarray is also limited to larger background noises due to hybridiza-

tion, such as cross-hybridization or non-ideal hybridization kinetics [Tu et al., 2002, Klebanov and Yakovlev, 2007].

1.1.4 RNA Sequencing (RNA-seq) of mRNA transcriptome

High-throughput sequencing technologies such as RNA-seq [Wang et al., 2009b] opens a new era for investigation on the mRNA transcriptome. Generally, in an RNA-seq experiment, the probed RNA molecules in the target transcriptome can be first synthesized into double stranded cDNAs, followed by a monitored process of fragmentation that cuts the full-length cDNAs into shorter pieces. A sample of the generated fragments usually with constrained length range (required by many sequencer) would be selected to construct a cDNA library for further sequencing. The output of the RNA-seq experiment is the single-end reads or paired-end reads, typically of length 100 200 bp, which are sampled from one end (single-end sequencing) or both ends (paired-end sequencing) of the size-selected fragments. See Figure 1.5. If paired-end sequencing is utilized, the original transcript fragments in the sample may be inferred according to the distribution of the mate-pair distances estimated from the data. Therefore, the produced RNA-seq reads are snapshots of subsequences of original mRNA molecules in the transcriptome.

The RNA-seq technique has several advantages over microarrays. First, by directly sequencing the cDNA fragments, RNA-seq allows the investigation on known transcripts and the exploration on novel ones. Second, It is capable of quantifying a larger dynamics range of expression levels, with absolute rather than relative values [Wang et al., 2009b]. Last, the hybridization issues seen with microarrays is

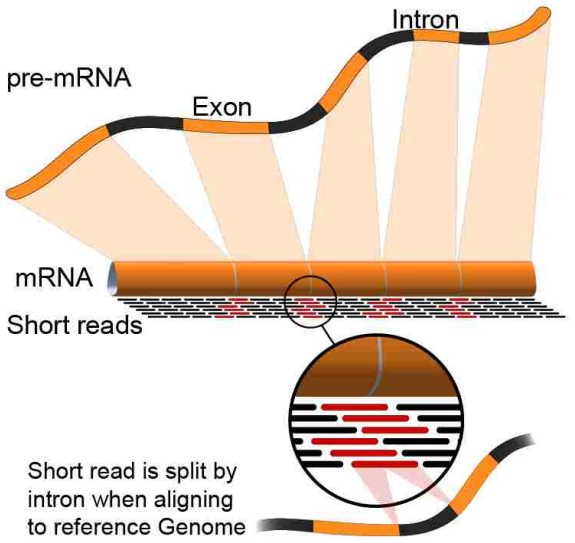


Figure 1.5: RNA-seq. Short RNA-seq reads are sequenced from mRNA transcripts in the transcriptome. Figure accommodated from `www.wikepedia.com`
.

eliminated in RNA-seq experiments. All these features make it suitable for transcript level reconstruction and quantification.

Although the application of next-generation sequencing or high-throughput sequencing technologies on transcriptome analysis has not been not long, just since year 2008 [Marguerat et al., 2008, Wang et al., 2009b], 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 new type of biological data has raised many computational and methodological challenges that excite the society of computational biology and bioinformatics, recognized by its extraordinary volume and computational difficulty. The sequence file of every single sample may take up to tens of Gigabytes in its binary format, consisting of tens or hundreds of millions of read records, requesting unprecedented challenges in issues such as data hosting, management, sharing, analyzing, and privacy control. On the

other hand, the sequencing capabilities of current next-generation sequencing protocols still limit the direct profile 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 interpretation of sequencing output in practice is often complicated by read length, sampling coverage, sampling errors and various types of sampling biases. Some sequencing protocols, such as platforms provided by 454 and Pacific Biosciences, may sequence reads up to thousands of nucleotides that may directly showcase the sequences of most transcripts and quantitate their abundance at the same time. Nonetheless the throughput and the quality are often highly limited, leading to insufficient coverage on the transcriptome. The RNA-seq experiments are mainstreamed by the short read sequencing protocols provided by Illumina. This protocol generate reads typically of a length less than 100 nucleotides, a length insufficient to identify the original transcript for most of the reads. Great ambiguity exists in the survey of transcript isoforms present in a sample and in the evaluation of their expression levels. Therefore, more and more approaches have been developed for the computational solutions that bridge short read sequences and biological findings. In the next section, we will review some of the primary computational challenges emerged in short-read sequencing.

1.2 Computational analysis using RNA-seq

Compared with early achievements which investigate the genome mainly at gene level, i.e. computational analysis is conducted on gene sequences, RNA-seq dives into a higher resolution. It allows us to look at alternative splicing events, gene fusion

and mutation/SNPs at mRNA transcript level, all of which are very important events and may potentially relate to diseases.

With all these benefits brought by RNA-seq, urgent needs escalate for bridging the gaps between the sequenced RNA-seq reads and the characterization of the transcriptome. However, the sequenced RNA-seq reads carry only partial information of original mRNA transcripts. Therefore, there exist many challenges for this task, such as: read mapping which tries to find the exact location on the genome where each RNA-seq reads may originate from, transcriptome assembly which aims at identifying the mRNA transcripts presented in the cell, transcript isoform expression estimation which quantifies the expression level of mRNA transcrips and *etc.* In this dissertation, we primarily focus on the transcriptome assembly and transcript quantification problems (quality and quantity assessment), which will reveal all characteristics of the mRNA transcriptome and is critical for downstream studies, like differential analysis between diseased and normal cells.

In this section, we will formally define the problem of transcriptome assembly and transcript quantification and show the challenges in solving these problems. First, we will briefly introduce RNA-seq read mapping.

RNA-seq reads alignment

The observed RNA-seq reads are sequences of 'A', 'C', 'G' or 'T', representing DNA bases. RNA-seq read alignment aims at locating the exact genomic coordinates on the genome where these reads are sampled from. This is achieved by aligning the reads to a reference genome. A reference genome can be considered as a assembled genome

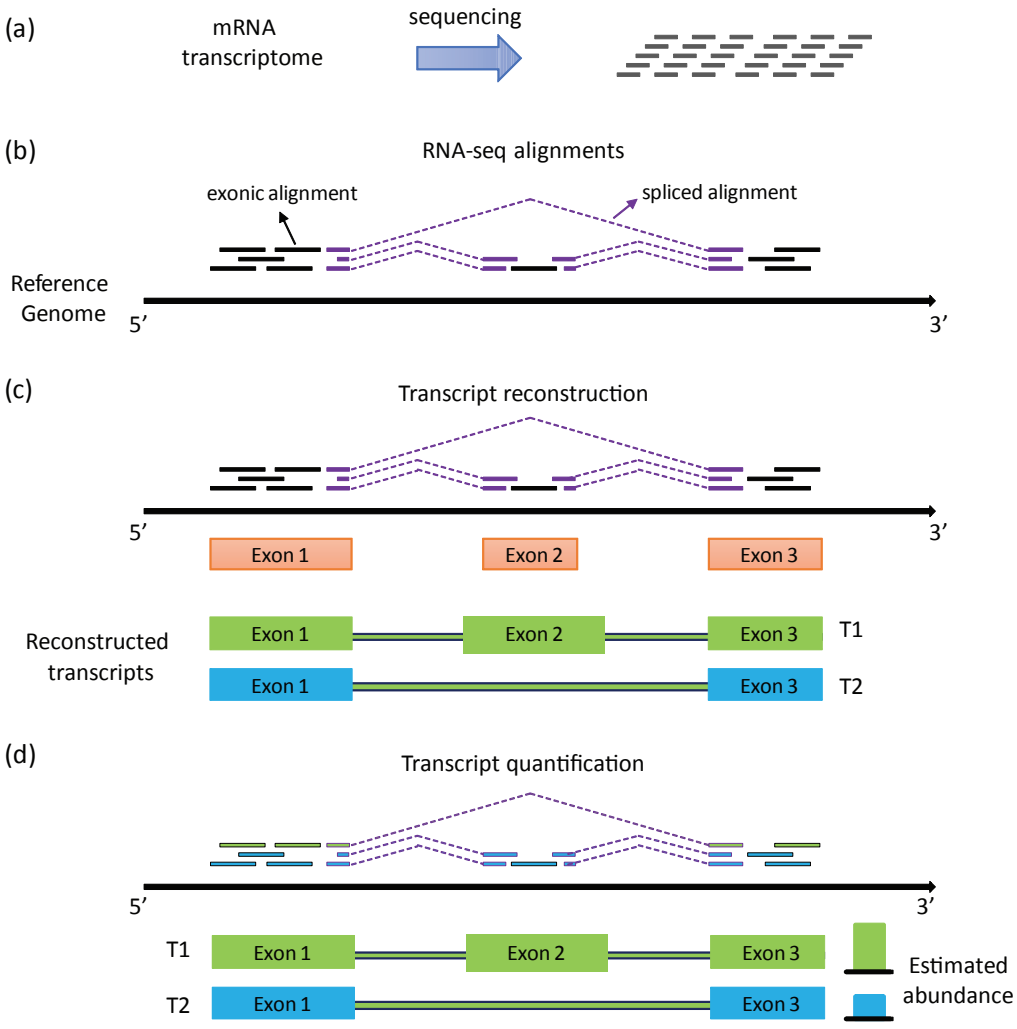


Figure 1.6: Typical computational analysis with RNA-seq data. (a) RNA-seq reads are sequenced from mRNA transcriptome. (b) RNA-seq short read alignment to the reference genome. (c) Transcriptome assembly. (d) Transcript quantification.

database which gathered from the sequencing of DNA from a number of donors. When a read is mapped as an entirety, it is referred as "exonic alignments"; otherwise, it is referred as "spliced alignments" which spans multiple exons and consequentially defines the splice sites of the splice junctions (Figure 1.6(b)). Here, the splice junctions are actually the intron region on the genome before alternative splicing happens.

Transcriptome assembly

Transcriptome assembly is a central problem of RNA-seq analysis, which aims at recapitulating the variety of transcript isoforms in an mRNA transcriptome from from the sequenced short reads. This procedure allows us to recover the genes and isoforms in existing database, as well as to detect novel ones. A typical RNA-seq protocol works by randomly fragmenting the mRNA transcripts followed by sequencing a sample of the total fragments. Therefore, the sequences of RNA-seq reads carry partial information of the original transcripts. After mapping the reads to the reference genome, the genomic coordinates where the reads are aligned will help reveal the exonic segments on the genome. The contiguous bases covered by read alignments constitute exons of a particular gene. While the splice junctions of the spliced alignments infer the intronic regions of that gene and indicate how the exons will be connected to form an isoform.

Figure 1.6(c) illustrates one simple example of transcript reconstruction. Three exons have been suggested by the RNA-seq read alignments along with three introns implied by the splice junctions. According to the observed evidence, two transcript isoforms can be reconstructed. One connects all three exons and the other skips the middle exon.

Transcript quantification

Recent studies have estimated that as many as 95% of all multi-exon genes are alternatively spliced, resulting in more than one transcript per gene [Pan et al., 2008b,

Wang et al., 2008b]. *Transcript quantification* determines the steady state levels of alternative transcripts within a sample, enabling the detection of differences in the expression of alternative transcripts under different conditions. Its application in detecting biomarkers between diseased and normal tissues can greatly impact biomedical research. Using RNA-seq data, the quantity of one transcript isoform expressed is usually measured by a statistic related to the number of reads falling on it. This statistic is calculated to approximate the number of mRNA molecule copies of such isoform. However, this task is not trivial. Since isoforms within a gene share sequences, sometimes it is difficult to assign one read to a specific isoform. For example, as showed in Figure 1.6(d), there are five exonic reads aligned on "Exon1". Solely based on this observation, we can hardly determine their origination. Through quantification, we may identify four reads (green) coming from T_1 and 10 reads (blue) coming from $T2$. Therefore, the relative abundance of these two isoforms are $1 : 2.5$.

Challenges in transcriptome assembly and quantification

Despite plenty of methods have been developed for solving the transcript reconstruction and quantification problems, they are still considered quite challenging. First, it is commonly observed that "the more the isoforms, the harder to predict" [Li et al., 2011b]. Intuitively, transcript isoforms from the same gene often overlap significantly. Limited by current sequencing technology, the length of RNA-seq read is insufficient (usually shorter than 200bp for a single-end read or about 500bp for a paired-end read). A short read may be mapped to more than one transcript isoform. Determining the presence and expression of individual transcripts from short read alignment,

therefore, can lead to an *unidentifiable* model, where no unique solution exists. Secondly, various sampling biases have been observed regularly in RNA-seq datasets as a result of library preparation protocols. These biases typically include position-specific bias [Bohnert and R., 2010, Li et al., 2010a, Roberts et al., 2011, Wu et al., 2011] such as 3' bias and transcription start and end biases, and sequence-specific bias [Li et al., 2010b, Roberts et al., 2011, Turro et al., 2011], where the read sampling in the transcriptome favors certain subsequences. How to compensate for these biases is an open problem. Finally, though closely related, transcriptome assembly and quantification are usually treated independently by existing methods. Typically, transcript reconstruction methods are first employed to produce a candidate set of isoforms, and quantification approaches may be further applied on the assembled set of isoforms or the reference database to estimate their abundance. However, this strategy may increase the risk of quantifying a false or incomplete set of transcripts. Moreover, it is biologically unlikely to expect all candidate transcripts for a given gene to be significantly expressed concurrently in a cell. Existing analytical approaches tend to assign positive expression values to every candidate transcript provided, thereby creating a situation in which large errors in abundance estimation can be computationally introduced for transcript isoforms that may, in reality, barely be expressed.

To address these challenges, in this dissertation, we focus on developing a novel and robust framework for comprehensive analysis of mRNA transcriptome both qualitatively and quantitatively, along with an accurate and consistent transcript abundance measure.

1.3 Thesis Statement

The aim of this dissertation is to develop computational methods for precise transcriptome analysis using RNA-seq data. Three closely related problems are studied: how to accurately estimate the abundance of the known transcript isoforms, how to simultaneously discover the presence and quantities of novel transcripts, and how to scale the transcriptome analysis to large-scale RNA-seq data. With prior knowledge of annotated gene/isoforms, a generalized linear model has been developed to solve the first problem which resolves the "unidentifiable" challenge in transcript quantification and effectively handles the sampling biases. For the purpose of detecting transcripts uncatalogued in existing database, a novel framework is designed that takes advantage of the biological interpretability of read relations and simultaneously infers the identities and quantities of the full-length gene isoforms residing in original cells. Empowered by the advent of large, complex clinical RNA-seq datasets, a systematic pipeline is built which aims to leverage information from massive samples and highlight meaningful transcription signals. All developed methods explore efficient solutions of recovering the characteristics of mRNA transcriptome both qualitatively by assessing the diversity of the mRNAs and quantitatively by estimating the abundance of the mRNAs from the RNA-seq read alignments. Meanwhile, they fully enable the translation from raw sequencing data to clinical insights and also provide valuable information differentiating functions of normal cells and diseased ones.

1.4 Contributions of this dissertation

In this dissertation, we have developed a series of computational methods targeting a comprehensive analysis of the mRNA transcriptome. All methods are data-driven. Salient information are mined and extracted from large-scale biological data, *i.e.*, raw RNA-seq reads. Experiments on both simulated and real RNA-seq datasets have demonstrated significantly improved sensitivity and specificity of our developed methods as compared to other state-of-the-art approaches.

Accurate transcript abundance estimation given reference annotation

Transcript quantification is performed to determine the steady state levels of all the alternative transcripts within a sample if a set of reference transcripts is provided. The reference transcripts could either from annotation database, or from various transcript assembly softwares. A robust model has been developed, named *MultiSplice*, which directly resolves three main challenges in the abundance estimation task: (1) ambiguity in solution; (2) bias in read sampling and (3) low-expression transcripts.

First, MultiSplice adopts a general linear model which not only includes information from single exons and splice junctions, but also leverages reads spanning multiple splice junctions to ameliorate unidentifiability. Second, all possible sampling biases are taken into account, like positional bias and sequence bias. The bias parameters are embedded into the general model. Lastly, to achieve reasonable sparsity. LASSO is utilized to solve the linear system in order to infer an accurate set of dominantly expressed transcripts.

Simultaneous transcript reconstruction and quantification

When the reference transcript database is incomplete or inaccurate, the transcriptome assembly is needed for producing a complete and correct set of transcript isoforms. An efficient and accurate algorithm has been proposed for simultaneous transcript reconstruction and quantification directly from RNA-seq paired-end read alignments.

We have developed a novel method named Astroid for simultaneous transcript reconstruction and quantification directly from RNA-seq paired-end read alignments. Recall that in a typical RNA-seq experiment, mRNA molecules in the sample are cleaved into fragments. Fragments with desired sizes are randomly selected and sequenced at one end (single-end sequencing) or both ends (paired-end sequencing). Using paired-end sequencing, the original transcript fragments in the sample may be inferred according to the distribution of the mate-pair distances estimated from the data. However, the distance between sampled fragments, which disconnects the fragments that belong to the same mRNA copy of a transcript, has barely been studied. Existing methods typically overlook the relation among the transcript fragments and assume independent sampling for the fragments. We instead propose to statistically model the distance between sampled transcript fragments, and to use this information to relate fragments and thread the observed reads into individual transcript copies. The read alignments are represented using vertices of a flow network, connected by edges that represent mate-pair distances and between-fragment distances. The likelihood of each edge is evaluated according to the distance distributions pre learned or specified distribution. A maximum likelihood set of transcript copies is

then reconstructed by solving a minimum-cost flow problem on the flow network. The number of copies for a transcript simultaneously provides a direct estimate for the transcript quantity. Lastly, we introduce a set of rules that clusters homogeneous vertices and edges and compresses the flow network. A compression parameter is defined to leverage the time and space complexity required by the flow network and the model accuracy.

Transcriptome analysis on large-scale data

The rapid development of sequencing technology allows us to sequence a sample or tissue at a much lower cost. For example, nowadays an RNA-seq experiment typically costs less than $\$1,000$, comparable to the cost of microarray. Recently, several projects have been launched which take advantage of this advancement and sequence tens of thousands of samples aiming at a comprehensive understanding of cell functioning as well as cell differentiation, such as TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium). The massive amount of data not only brings more power for expanding our knowledge of human genome, but also introduces great challenges. First, high volume of data means high demand of computing and storage resources. Second, large-scale data from heterogeneous samples/tissues incurs ambiguity of a overall analysis: the uncertainty will be extremely amplified when examining all assemblies simultaneously. Therefore, the joint analysis of hundreds of transcriptome is not simply a trivial extension from existing methods.

Driven by the desire of finding biological signatures from TCGA breast cancer projects (819 samples included), we systematically investigate the current standard

pipelines for processing large-scale RNA-seq datasets. They can be divided into two categories: reference transcriptome guided and reference transcriptome independent. The first one solely quantifies the annotated transcripts and the latter one starts by reconstructing a transcript set then estimates their abundances. Both strategies treat each individual sample independently which involves heuristic and potential filtering on every sample. Furthermore, their performances are limited by the difficulties of assembling and quantifying full-length transcripts on massive datasets. Alternatively, we have developed an *ab initio* workflow which establishes a joint analysis model that summarizes all samples with a single splice graph without the knowledge of gene/isoform annotations. In stead of per-sample analysis, information is pooled together for detection of aberrant alternative splicing markers. To our best knowledge, this is the first method directly targeting the above challenges and dedicated to large-scale transcriptome analysis.

Summary

If we consider the sequencing process as fragmenting the mRNA transcriptome into millions of smaller pieces, each of which corresponds to a short RNA-seq reads. The effort of this dissertation is like seeking solutions of piecing the puzzle together to recapitulate the original picture. The problem itself is quite challenging due to the large quantity of puzzle pieces. It can be further complicated by real applications, such as: sampling biases introduced during sequencing procedure, analogous to the circumstance that pieces from another puzzle set are mixed in; the need of comparing a large set of similar but slightly different puzzles, analogous to the comparison of hun-

dreds of transcriptome; and *etc*. The methods developed in this dissertation directly address the computational challenges in "solving the puzzle" and have demonstrated superior effectiveness and efficiency in transcriptome analysis.

Chapter 2 Related work in RNA-seq-based transcriptome analyses

In this chapter, we will summarize the general analysis of mRNA transcriptome using RNA-seq data including the computational challenges and the current solutions.

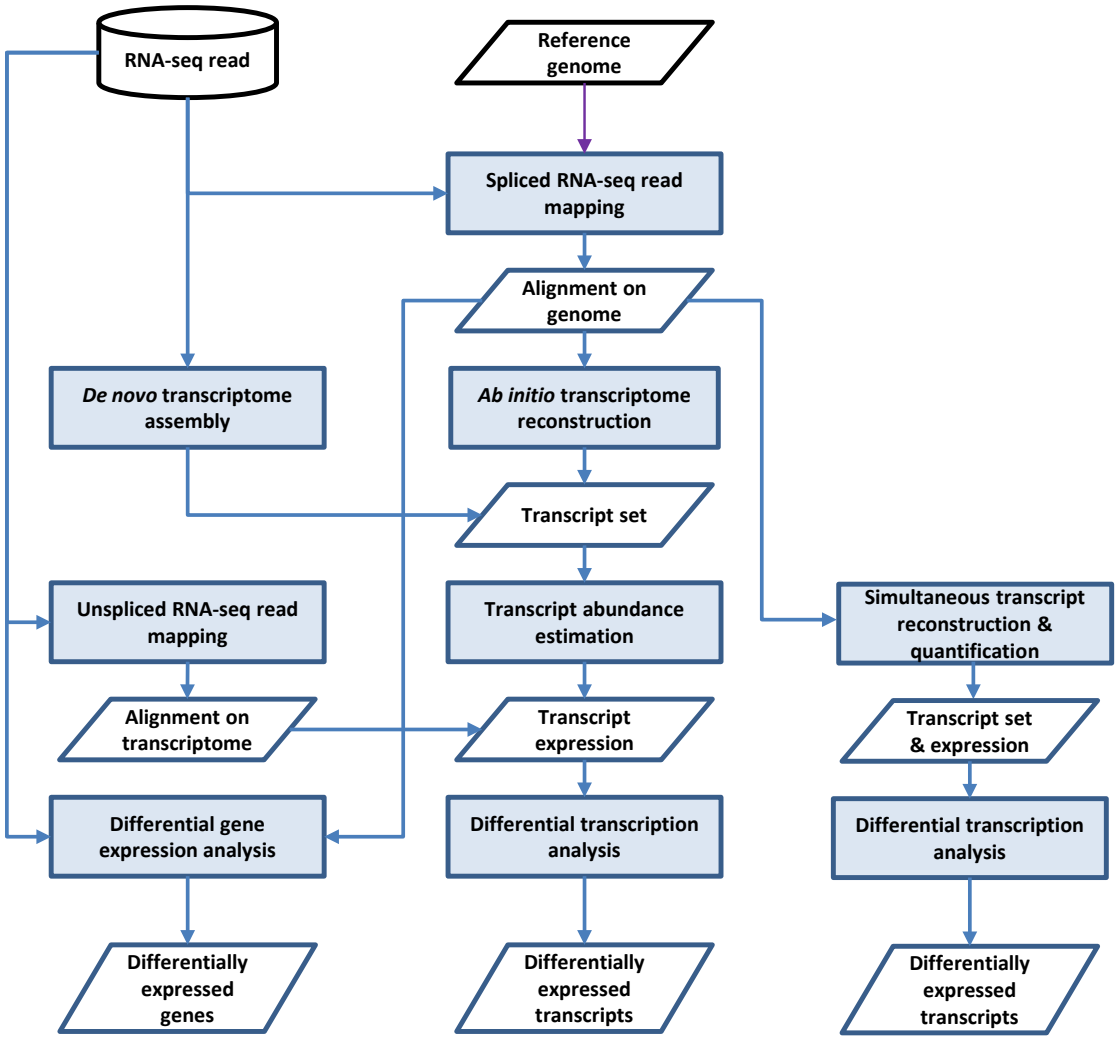


Figure 2.1: The typical workflows in the transcriptome studies using RNA-seq technologies.

Figure 2.1 illustrates a typical workflow in the transcriptome studies. Although the high-throughput RNA-seq reads provides an unprecedented opportunity to pre-

cisely profile the mRNA transcriptome of a specific cell population, the observed RNA-seq reads are still local pieces of the original transcripts, bringing great ambiguity to the effort of portraying the transcriptome. How to reconstruct the original transcriptome with the RNA-seq reads remains a challenging problem. Generally, the current methodologies of addressing this problem can be divided into two categories: *reference-guided* and *reference-independent* referring to whether or not the analysis is guided by the reference genome or transcriptome. The reference-independent methods are greatly useful when a reference genome is not available or when individual modifications to the reference genome is significant. The representative approaches are Trinity [Grabherr et al., 2011] and Trans-ABySS [Birol, 2009], which are known as the *de novo* assembly. They assemble mRNA transcripts solely based on nucleotides sequenced in RNA-seq reads without the guidance of any reference, which may followed by downstream analysis (Figure 2.1). When a reference genome/transcriptome is accessible, as for human for example, the other category, usually first align the reads to the reference genome [Wang et al., 2010a, Trapnell et al., 2009b] or transcriptome [Langmead et al., 2009a]. Transcripts can be reconstructed and quantified according to the genomic coordinates or the mapped reads. Compared with *de novo* methods, it is computational efficient regarding the time cost and memory usage, and most importantly, it has been demonstrated to have higher sensitivity and specificity in the subsequent assembly step [Garber et al., 2011b]. Here, we limit the discussion of this dissertation in the scope of reference-guided methods. In the following sections we will briefly introduce the existing work for reference-guided transcriptome studies and their problems.

2.1 RNA-seq read alignments

There exist two threads of work of read mapping approaches: the unspliced aligners and the spliced aligners [Garber et al., 2011a].

The unspliced aligners usually rely on the access to the exact or similar reference transcriptome of a specified species. Reads are mapped to a reference without allowing any large gaps. The representative work include but not limited to MAQ [Li et al., 2008], SHRiMP [Rumble et al., 2009], ELAND [Cox, 2007], Novoalign [Hercus], Stampy [Lunter and Goodson, 2011], Bowtie [Langmead et al., 2009b], BWA [Li and Durbin, 2009], Bowtie 2 [Langmead and Salzberg, 2012] and SNAP [Zaharia et al., 2011].

However, the nature of unspliced aligners make them of limited use since they can only identify known exons and splice junctions. Alternatively, the spliced aligners map the reads to the genome where reads can span multiple exonic regions separated by introns on the genome. This kind of methods allow the detection of novel exons, junctions, and therefore novel transcript isoforms, making it more suitable for a comprehensive analysis of the mRNA transcriptome. Several methods are developed in this category, including: TopHat [Trapnell et al., 2009b], SpliceMap [Au et al., 2010], MapSplice [Wang et al., 2010b], GSNAP [Wu and Nacu, 2010], STAR [Dobin et al., 2013] and etc.

By first mapping the reads to the reference genome/transcriptome, we obtain the genomic coordinates of all possible exon and junction boundaries while build a foundation for recapitulating the mRNA transcriptome.

2.2 *Ab initio* transcript reconstruction

A handful of computational methods for transcript reconstruction have been developed to bridge the gap from the sequenced short read alignments to the identity of the original transcripts [Trapnell et al., 2010b, Guttman et al., 2010, Li et al., 2011b, Huang et al., 2012].

A common simplification in existing transcript reconstruction approaches is to reconstruct transcripts from a small set of features extracted from the reads. Most approaches, for example, Scripture [Guttman et al., 2010] and IsoLasso [Li et al., 2011b], build a splice graph [Heber et al., 2002, Hu et al., 2012] in which nodes stand for the exons and the edges stand for the splice junctions. Each path on the splice graph represents a possible transcript. Cufflinks [Trapnell et al., 2010b] summarizes the read alignments with a partial order list, based on the mutual compatibility whether two reads may be explained by one transcript. A set of expressed transcripts is then solved on the splice graph or the partial order list, typically combined with transcript-level heuristics or shrinkage such as maximum parsimony [Trapnell et al., 2010b], maximum sensitivity [Guttman et al., 2010] and Lasso [Li et al., 2011b] (Figure 2.2).

However, the simplification relying on extracted features that current methods made ignore the relation among the reads – as long as the features extracted from the reads, or the probabilities of the read being sampled from each transcript, or the collective statistics do no change, how the reads are distributed cannot provide additional information.

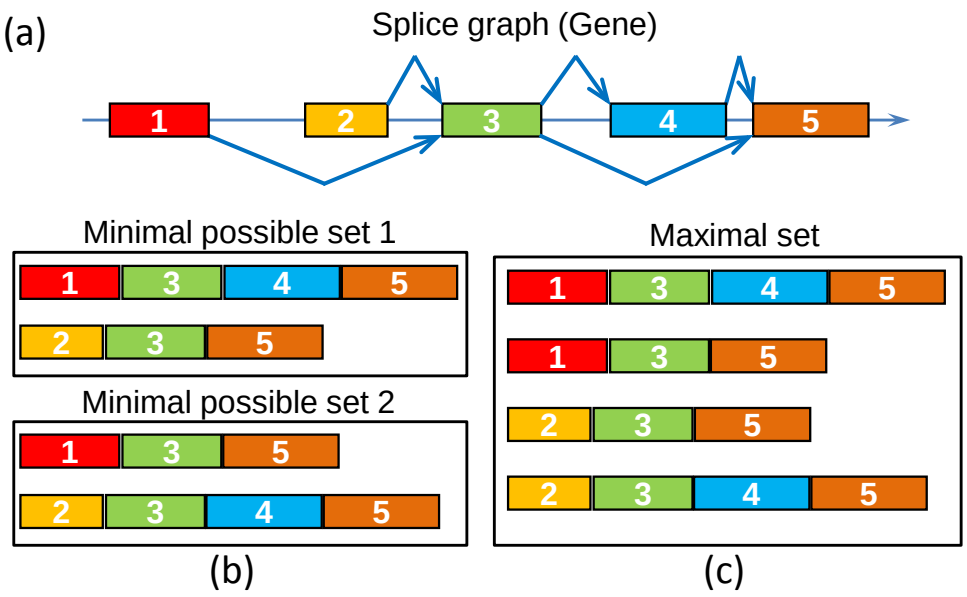


Figure 2.2: (a) Splice graph built from RNA-seq read alignments. (b) and (c) show two different strategies of transcript assembly. (b) refers to *maximal parsimony* which gives rise to only two isoform transcripts that can explain all reads. (c) refers to *maximum sensitivity* that generates all possible transcript isoforms from the splice graph.

Moreover, since they typically work from the exons and the splice junctions revealed from the RNA-seq read alignment. One central difficulty of transcript reconstruction becomes solving the combinatorial ambiguity in linking splice isoforms of different alternative splicing events into full transcripts. Heuristics are what the existing assembly methods usually rely on. For example, on the basis of maximum parsimony, Cufflinks Trapnell et al. [2010b] will choose the minimum set of transcripts that can explain the observed fragments. 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. [2011b], 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. Nonetheless, the intention of these heuristics focus on how to select one optimal combination of exons and splice junctions which is the secondary structure inferred from read or fragment alignments, while ignoring the linkage relationship among the transcript fragments which would reveals the original mRNA molecules more directly.

2.3 Transcript abundance estimation

The problem of transcript quantification is often treated separately from transcript assembly. A common simplification is to assume independent and random sampling of reads. This assumption allows processing each read individually with a same model to calculate the probability that a read is sampled from a transcript [Trapnell et al., 2010b, Li and Dewey, 2011, Nicolae et al., 2011]. Alternatively, this assumption allows efficient inference with only a few collective statistics, such as the number of reads mapped to each exon [Jiang and Wong, 2009, Huang et al., 2012, Bohnert and R., 2010].

However, these simplifications ignore the relation among the reads – as long as the features extracted from the reads, or the probabilities of the read being sampled from each transcript, or the collective statistics do no change, how the reads are distributed cannot provide additional information. For example, the two sets of reads in

Figure 2.3a suggest the same splice graph hence contain the same information regarding possible transcripts. According to maximum parsimony, for instance, the two sets of reads may be explained by the same two transcripts. However, the distribution of the reads may suggest two different sets of transcripts. Figure 2.3b, shows two read distributions that contain the same number of reads but have drastically different coverage profiles. Collectively, the two loci will get the same FPKM Trapnell et al. [2010b], Li et al. [2011b], *i.e.*, the Fragments Per Kilobase of transcript per Million mapped reads. However, the relative location of the reads in set 2' may suggest an additional transcription termination that distinguishes two transcripts of different expression levels. This also demonstrates the advantage of performing transcript reconstruction and quantification simultaneously, as alleviating the risk of estimating transcript abundance on the basis of an incorrect set of transcripts.

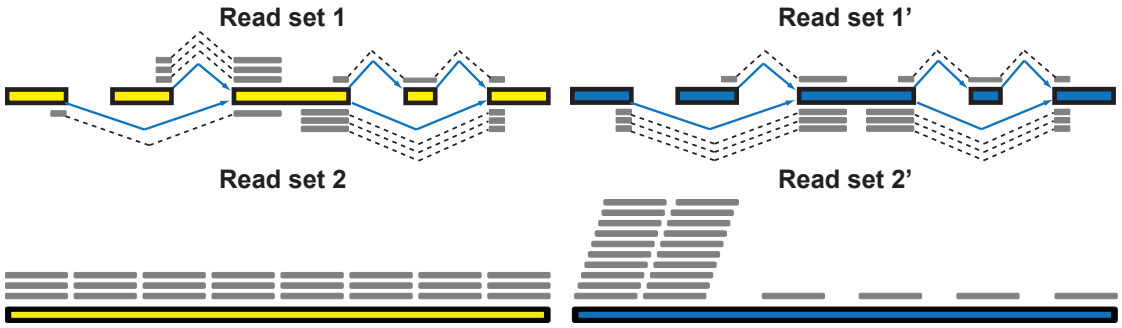


Figure 2.3: The read sets 1 and 1'. Two read sets have the same number of reads. The read sets 2 and 2', Two read sets have the same number of reads. Different read distributions may suggest different set of transcripts.

Another problem is the current measurement for transcript quantity based on read count may be skewed in practice. The typical unit FPKM requires that the number of fragments sampled from each transcript is strictly proportional to the length and the number of molecule copies of the transcript. However, this correlation may become

poor especially for short transcripts (*e.g.* $\sim 14\%$ of transcripts in human genome are less than 1000nt, according to UCSC GAF 2.0 annotation), because these transcripts may be fragmented into smaller pieces that will not be sequenced due to size selection in RNA-seq James [2011]. The presumed correlation may further weakened by sampling biases (*e.g.* GC-content biases and positional biases Bohnert and R. [2010], Roberts et al. [2011]) and read mapping errors. Furthermore, FPKM normalizes relative expression according to the total number of mapped reads in one sample, which may not reflect the true library size and may bias comparison of transcript expression across samples Wagner et al. [2012], Dillies et al.. Another measure TPM [Li and Dewey, 2011, Wagner et al., 2012], *i.e.*, Transcripts Per Million, resolves this inconsistency problem. It approximates the transcript number by normalizing the cumulative per base read coverage by the isoform length. The library size is estimated by summing up the estimated abundance of all isoforms accounting for the total number of transcripts in the transcriptome. However, it is unclear how well the per base coverage in TPM can approximate the true abundance of one isoform because it is impossible that all observed fragments can be tightly arranged one after the other making every single base of the isoform covered by the read.

Chapter 3 A Robust Method for Transcript Quantification with RNA-seq Data

In this chapter, we introduce a robust method for transcript-level quantification. Transcript isoforms can differ not only in exons alternatively included or excluded but also in where two or more exons are connected together. In RNA-seq data, this information is typically implied by the spliced reads, i.e., the reads that cross one or more splice junctions. We have developed a general linear model for transcript quantification that leverages discriminative features in spliced reads to ameliorate the issue of identifiability and simultaneously corrects the sampling bias. Our contribution of this method is three-fold: (1) We explicitly identify *MultiSplice*, a novel structural feature consisting of a contiguous set of exons that are expected to be spanned by the RNA-seq reads or transcript fragments of a given length. The MultiSplice, which includes single splice junctions as a special case, is used in two ways: its presence in the sample will infer the host transcript while its absence may reject it. MultiSplices are more powerful than single exons in disambiguating transcript isoforms, making more transcript quantification problems identifiable with long or paired-end reads; (2) We set up a linear system which minimizes the summed relative squared errors regarding the ratio of the expected expression against the observed expression across all structure features along a gene while taking into account various bias effects; (3) We develop an iterative minimization algorithm in combination with LASSO [Tibshirani, 1996] to resolve the aforementioned linear system in order to achieve the most accurate set

of dominantly expressed transcripts while simultaneously correcting biases.

We have demonstrated the efficacy of our methods on both simulated RNA-seq datasets and real RNA-seq data: (1) We conducted the first study to investigate the question: what is the maximum read length needed in order to disambiguate all possible transcript isoforms in transcriptomes from different species; (2) We compared the proposed method with several state-of-the-art methods including Cufflinks, RSEM, the Poisson model, and the ExonOnly model. Our results using simulated data from the human mRNA transcriptome demonstrated superior performance of the proposed method in most cases. When applied to 8 RNA-seq datasets from two breast cancer cell lines (MCF-7 and SUM-102), the quantification obtained from MultiSplice demonstrated good consistency within technical replicates from each transcriptome-wide assessment and substantial differences between the two biological groups (cell lines) in a small percentage of genes.

3.1 Method

In this section, we propose a method designated *MultiSplice*, for mRNA isoform quantification. We first define the observed features used in the MultiSplice model and the statistics collected. Then, we derive a general linear model to relate transcript level estimate to the observed expression on every feature.

Preliminaries. For a gene g, we use $\mathcal{E}_g$ to denote the set of exonic segments [Jiang and Wong, 2009, Li et al., 2011b] in g, which are disjoint genomic intervals on the genome that can be included in a transcript in its entirety. We use $\mathcal{T}_g$ to denote the set of mRNA isoforms transcribed from g. These mRNAs can be a set of annotated

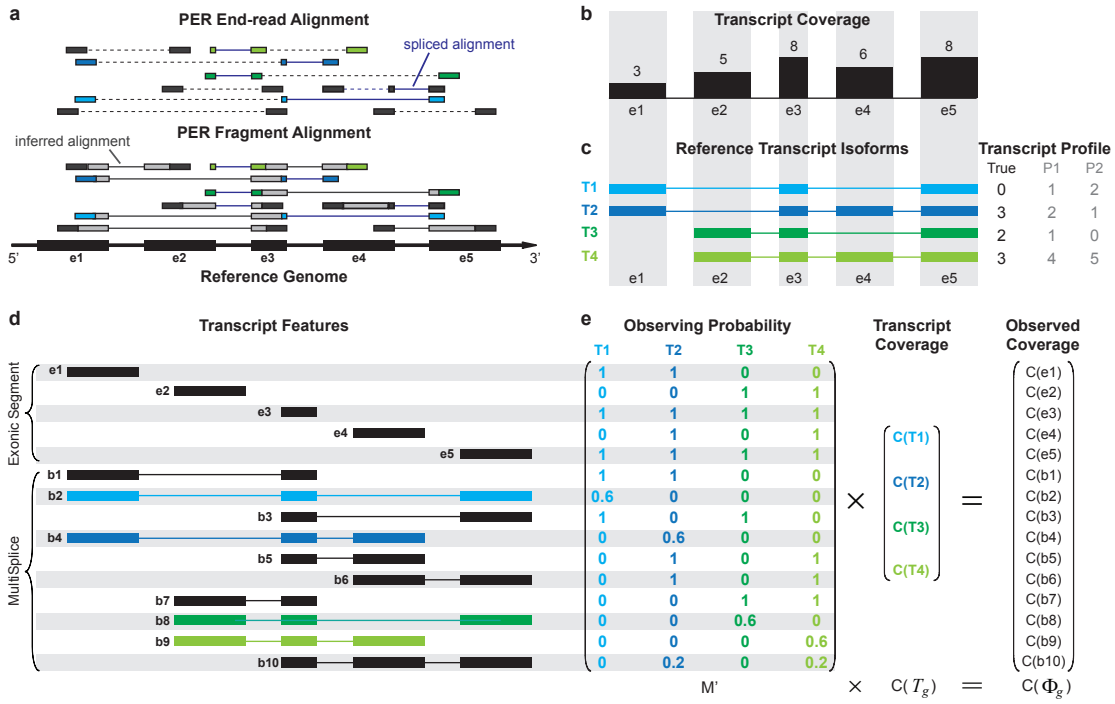


Figure 3.1: Overview of the MultiSplice model. **a.** Sequenced RNA-seq short-reads are first mapped to the reference genome using an RNA-seq read aligner such as MapSplice [Wang et al., 2010a]. In the presence of paired-end reads, MapPER [Hu et al., 2010] can be applied to find *PER fragment alignments* for the entire transcript fragment based on the distribution of insert size. **b.** Observed coverage on each exonic segment. **c.** Four transcripts originate from the alternative start and exon skipping events. Provided with these transcripts, abundance estimates would be unidentifiable for methods that only use coverage on exonic segments. Both transcript profiles P_1 and P_2, for instance, can explain the observed read coverage on each exon, but deviate from the true transcript expression profile. **d.** MultiSplices that connect multiple exonic segments in a transcript. **e.** A linear model can be set up where the expected coverage on every exonic or MultiSplice feature approximates its observed coverage. The transcript expression is solved as the one that minimizes the sum of squared relative error.

transcripts retrieved from a database such as Ensembl [Ens] or Refseq [Ref]. A transcript $t \in \mathcal{T}_g$ is defined by a sequence of exon segments, $t = e_1^t e_2^t \cdots e_{n_t}^t$, where $e \in \mathcal{E}_g$ and n_t denotes the number of exonic segments in the transcript t. The length of each exonic segment e is defined as the number of nucleotides in the exonic segment, denoted as $l(e)$. Hence, the length for every transcript is $l(t) = \sum_{i=1}^{n_t} l(e_i^t)$.

3.1.1 MultiSplice

In a typical RNA-seq dataset, a significant percentage of the read alignments are spliced alignments that connect more than one exon. With paired-end reads, the transcript fragment where its two ends are sampled can be inferred based on the distribution of the insert size [Roberts et al., 2011]. Transcript fragments are typically between 200bp and 300bp, making them more likely to cross multiple exons, indicating these exons are present together in one transcript. This information can be crucial in distinguishing alternative transcript isoforms. However, they are often ignored in current computational approaches.

In this subsection, we consider a sequence of adjacent exons in an mRNA transcript covered by transcript fragments. These structural features are the basis of *MultiSplice*. For generality, we assume that the RNA-seq reads are sampled from transcript fragments whose lengths follow a given distribution F_{fr} with probability density function f_{fr}. For example, the fragment length distribution F_{fr} is often modeled as a normal distribution with mean and variance learned from the genomic alignment of the RNA-seq reads. We also assume the maximum fragment length is l_{fr}.

Definition Let $b = e_i^t e_{i+1}^t \cdots e_{i+n_b}^t$ be a substring of a transcript sequence $t = e_1^t e_2^t \cdots e_{n_t}^t$, $n_b \geq 1$ and $i + n_b \leq n_t$. Then b is a MultiSplice in t if and only if

$$\sum_{q=1}^{n_b-1} l(e_{i+q}) \leq l_{fr} - 2. \tag{3.1}$$

The condition in Equation 3.1 guarantees that a MultiSplice b connects $n_b + 1$ adja-

cent exons with at least 1 base landed on the 5' most exon e_i^t and the 3' most exon $e_{i+n_b}^t$. We use $\mathcal{B}_g$ to denote the set of all MultiSplices in gene g. From the definition, the set of MultiSplices vary according to the fragment length l_{fr}. The longer the fragments, the more MultiSplices are expected to function as structural features, and the higher power in disentangling highly similar alternative isoforms.

In Figure 3.1, for example, assume the maximum fragment length is $l_{fr} = 300bp$ with the expected fragment length of $250bp$ and the exonic segments of this gene have lengths of $l(e_1) = 200bp, l(e_2) = 200bp, l(e_3) = 100bp, l(e_4) = 200bp, l(e_5) = 200bp$. In reference transcript $T_1 = e_1e_3e_5$, $b_2 = e_1e_3e_5$ is a substring of T_1, and we have $l(e_3) = 100bp < 300bp = l_{fr}$ which allows a fragment to cover b_2. Therefore, b_2 is a MultiSplice feature of the gene. Combining MultiSplices from all the reference transcripts, b_1, b_3, b_5, b_6, and b_7 are MultiSplices consisting of a single splice junction, b_2, b_4, b_8, b_9, and b_{10} are MultiSplices consisting of two splice junctions.

3.1.2 Expected coverage and observed coverage

Given the gene g and a transcript $t \in \mathcal{T}_g$, let c_i be the number of transcript fragments covering the ith nucleotide of t. We define the coverage on t as the averaged number of transcripts covering each base in the transcript, $C(t) = \frac{1}{l(t)} \sum_{i=1}^{l(t)} c_i$. Then $C(t)$ is an estimator for the quantity of t in the sample, which provides a direct measure for the expression level of t. In our model, $C(t)$ is the unknown variable. The feature space that can be observed from the given RNA-seq sample is the union of all exonic segments and MultiSplices of the gene, $\Phi_g = \mathcal{E}_g \cup \mathcal{B}_g$. We aim at resolving the transcript expressions that minimize the difference between the observed expression

and the expected expression of every feature.

The *observed* coverage on an exonic segment $e \in \mathcal{E}_g$ is defined as $C(e) = \frac{1}{l(e)} \sum_{i=1}^{l(e)} c_i$, where c_i is the number of reads covering the ith nucleotide in e. The read coverage $C(e)$ provides an estimator for the number of transcript copies that flow through the exonic segment e assuming uniform sampling. For a MultiSplice $b \in \mathcal{B}_g$, we use $C(b)$ to denote the read coverage on b defined as the number of transcript fragments that include b.

For every $\phi \in \Phi_g$ and every transcript $t \in \mathcal{T}_g$, the expected coverage of feature ϕ from t can be expressed as a function of the transcript quantity $C(t)$, i.e., $E[C(\phi|t)] = m(\phi, t)C(t)$, where $m(\phi, t)$ contains the probability of observing ϕ in t assuming uniform sampling. Next, we define the *expected* coverage on exonic segments and MultiSplice respectively.

For an exonic segment e in t, assuming N_t fragments were sampled from t, the number of fragments falling in e then follows a binomial distribution with parameters N_t and $p(e|t)$, where $p(e|t) = \frac{l(e)}{l(t)}$ denotes the probability that a fragment sampled from t originated from e. Therefore, the expected number of reads on e from t is $E[N_{e|t}] = N_t p(e|t)$. Let $fr_1, fr_2, \cdots, fr_{N_t}$ be the fragments sampled on t, the expected fragment coverage on t is $E[C(t)] = E[\frac{\sum_{i=1}^{N_t} l(fr_i)}{l(t)}] = \frac{N_t E[l(fr)]}{l(t)}$, where $E[l(fr)]$ is the expected fragment length. On the other hand, the expected fragment coverage on e contributed by t is calculated as $E[C(e|t)] = E[E[C(e|t)|N_{e|t}]] = E[\frac{N_{e|t} E[l(fr)]}{l(e)}] = \frac{E[N_{e|t}]E[l(fr)]}{l(e)} = \frac{N_t p(e|t) E[l(fr)]}{l(e)}$. Since $p(e|t) = \frac{l(e)}{l(t)}$, $\frac{p(e|t)}{l(e)} = \frac{1}{l(t)}$. Therefore, we could get $E[C(e|t)] = \frac{N_t E[l(fr)]}{l(t)}$, which means the expected fragment coverage on e contributed by t equals the expected fragment coverage of t, which concludes that the probability

of observing e in t is 1: $m(e,t) = 1$.

For a MultiSplice $b = e_i^t e_{i+1}^t \cdots e_{i+n_b}^t$, we are interested in the number of fragments containing it. Should a transcript fragment fr cover b, fr must start no later than the 3' end boundary of the 5' most exonic segment e_i^t and have at least 1 base landed on the 3' most exonic segment $e_{i+n_b}^t$. Therefore, there exists a window $w(b)$ before the 3' end of e_i^t with length $l(w(b)) = l(fr) - \sum_{q=1}^{n_b-1} l(e_{i+q}) - 1$, where b can be covered by the transcript fragment fr. The probability that fr covers b in transcript t is hence $p(b|t) = \frac{l(fr)-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{l(t)}$. Equivalent to the expected number of fragments from t that contain b, the expected fragment coverage on b from t is $E[C(b|t)] = E[N_{b|t}] = E[N_t p(b|t)] = N_t \frac{E[l(fr)]-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{l(t)}$. Since $E[C(b|t)] = m(b,t)C(t)$, the probability that the MultiSplice b is observed within transcript t is $m(b,t) = \frac{E[C(b|t)]}{C(t)}$. Recall that $C(t) = \frac{N_t E[l(fr)]}{l(t)}$, therefore, $m(b,t) = \frac{E[l(fr)]-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{E[l(fr)]}$. In Figure 3.1, for example, $m(b_2, T_1) = \frac{E[l(fr)]-l(e_3)-1}{E[l(fr)]} = \frac{250-100-1}{250} = 0.6$.

In summary, the probability that a feature ϕ contained in a uniformly sampled transcript fragment f_r is:

$$m(\phi,t) = \begin{cases} 1 & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{E}_G \\ \frac{E[l(f_r)]-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{E[l(f_r)]} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{B}_G \\ 0 & \text{if } \phi \not\subset t. \end{cases} \tag{3.2}$$

with $\phi \subset t$ standing for that ϕ is contained in t.

3.1.3 A generalized linear model for transcript quantification

We construct a matrix $\mathbf{M}' \in \Re^{|\Phi_g| \times |\mathcal{T}_g|}$ to represent the structure of the transcripts, whose entry on the row of ϕ and the column of t corresponds to the probability of

observing feature ϕ from transcript t, $\mathbf{M}'(\phi, t) = m(\phi, t)$. The linear model is set up for every feature $\phi \in \Phi_g$ by equating the observed coverage on ϕ to the expected coverage from all transcripts:

$$C(\phi) = \sum_{t \in \mathcal{T}_g} \mathbf{M}'(\phi, t) C(t) + \epsilon_\phi, \text{for any } \phi \in \Phi_g. \tag{3.3}$$

Here $C(t) \geq 0$ for every $t \in \mathcal{T}_G$, ϵ_ϕ is the error term for feature ϕ in transcript t.

Lemma 3.1.1 *The MultiSplice model for transcript quantification is identifiable if the rank of M' is no less than the number of transcripts $|\mathcal{T}_g|$.*

Lemma 3.1.1 directly follows the the Rouché-Capelli theorem [Horn and Johnson, 1990].

3.2 Bias correction

Under uniform sampling, the sampling probability is the same at every nucleotide of a transcript. The observed coverage on ϕ is unbiased for the expected coverage on t. In this case, the bias coefficient $\sigma(\phi, t)$ is set to 1 for all transcripts and features. However, sampling bias is often introduced in RNA-seq sample preparation protocols and has been demonstrated to have significant effects in RNA-seq analysis [Kozarewa et al., 2009, Wang et al., 2009a]. Therefore, we discuss in the following subsections how MultiSplice corrects various sampling bias via learning of the bias coefficients and simultaneously solves the linear model for transcript coverage $C(t)$ of every transcript t.

Figure 3.5(a-e) shows how various types of sampling bias alter the sampling probability and hence the coverage. Two types of sampling bias are commonly observed in

RNA-seq data, namely, the position-specific bias and the sequence-specific bias [Bohnert and R., 2010, Roberts et al., 2011, Olejniczak et al., 2010, Srivastava and Chen, 2010]. In our model, sampling bias may affect the sampling probability of both the exonic segments and MultiSplices. Therefore, we calculate the bias coefficient $\sigma(\phi, t)$ for every feature $\phi \in \Phi_g$ and every transcript t so that $E[C(\phi|t)] = \sigma(\phi, t)m(\phi, t)C(t)$. Next, we introduce each independent bias individually.

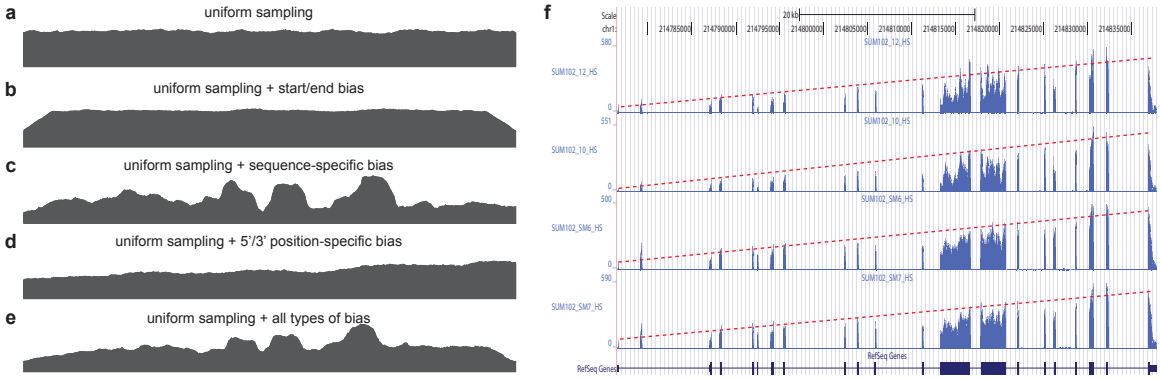


Figure 3.2: Sampling bias present in the RNA-seq data. **a**. RNA-seq read coverage under uniform sampling. **b**. RNA-seq read coverage under uniform sampling with transcript start/end bias. **c**. RNA-seq read coverage under uniform sampling with sequence-specific bias. **d**. RNA-seq read coverage under uniform sampling with 5'/3' position-specific bias. **e**. RNA-seq read coverage under uniform sampling with all aforementioned types of bias. **f**. Sampling bias on gene CENPF in the breast cancer dataset used in Section 6. Please note that the second peak in the coverage plot is not an exon in CENPF. The observed coverage on each exon decreases almost linearly from the 3' end to the 5' end. The coverage also drops at the bases near the end of the gene. The non-uniformity in the two middle large exons is likely to be due to the sequence-specific sampling bias.

3.2.1 Sequence-specific bias.

The sequence-specific bias refers to the perturbation of sampling probability related to certain sequences at the beginning or end of transcript fragments [Roberts et al., 2011, Li et al., 2010b]. The characteristic of this type of bias in the given RNA-seq

sample can be learned in advance by examining the relationship between GC content and the observed coverage on single-isoform genes. To derive the sequence-specific bias at an arbitrary exonic position, we look into 8bp upstream to the 5' start to 11bp downstream according to [Roberts et al., 2011]. A Markov chain is constructed to model the effect on the sampling probability at the position from the sequence of surrounding nucleotides. Then we use an approach based on the probabilistic suffix tree [Bejerano, 2004] to learn the sequence-specific bias coefficient $\alpha(t, i)$ for ith nucleotide in transcript t.

3.2.2 Transcript start/end bias.

Sampling near transcript start site or transcript end site is often insufficient. The read coverage in these regions is typically lower than expected because the positions where a sampled read can cover are restricted by the transcript boundaries. The bias coefficient for start/end bias at the ith nucleotide in transcript t is written as:

$$\beta(t,i) = \begin{cases} i/E[l(fr)] & \text{if } i < E[l(fr)] \\ 1 & \text{if } E[l(fr)] \leq i \leq l(t) - E[l(fr)] \\ (l(t)-i)/E[l(fr)] & \text{if } i > l(t) - E[l(fr)]. \end{cases}$$

3.2.3 5'/3' position-specific bias.

Position-specific bias refers to the alteration on sampling probability according to position in the transcript. For example, nucleotides to the 3' end of the transcript have higher probability to be sampled in Figure 3.5(f). Here we model the position-specific bias coefficient as a linear function, $\gamma(t, i) = \gamma_1^t \cdot i + \gamma_0^t$. The intercept γ_0^t gives the bias coefficient at the 5' transcript start site. The slope γ_1^t measures the extent of the bias: a positive γ_1^t indicates that 3' transcript end site has higher sampling

probability than the start site; a zero γ_1^t indicates no positional bias in the transcript t.

3.2.4 Combined bias model.

Assuming the above three types of bias have independent effect on read sampling, we derive the bias coefficient at ith nucleotide in transcript t as $\sigma(t,i) = \alpha(t,i) \cdot \beta(t,i) \cdot \gamma(t,i)$. The bias coefficient of an exonic segment $e \in \mathcal{E}_g$ is then the averaged bias coefficient on all positions in the exonic segment e, and the bias coefficient of a MultiSplice $b \in \mathcal{B}_g$ is the averaged bias coefficient on all positions in its sampling window $w(b)$. In summary, the bias coefficient for a MultiSplice feature $\phi \in \Phi_g$ in transcript t is

$$\sigma(\phi,t) = \begin{cases} \frac{\sum_{i\in\phi}\sigma(t,i)}{l(\phi)} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{E}_g \\ \frac{\sum_{i\in w\phi}\sigma(t,i)}{E[l(w(\phi))]} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{B}_g \\ 0 & \text{if } \phi \not\subset t. \end{cases} \tag{3.4}$$

3.3 Solving the general linear models with bias correction

Conventionally, we are interested in the set of transcript expressions that minimize the sum of squared errors, the absolute residuals between the expected coverage and the observed coverage. This solution is relatively sensitive to unexpected sampling noise which often occurs in real RNA-seq samples and may lead to a highly unstable extrapolation when the expression of the alternative splicing events discriminating the transcripts is notably lower than the average level of gene expression. Therefore, we define the sum of squared relative errors (SSRE), which measures the relative

residual regarding the ratio of the expected coverage against the observed coverage.

$$SSRE = \sum_{\phi \in \Phi_g} \left(\frac{\sum_{t \in \mathcal{T}_g} \sigma(\phi, t) \mathbf{M}'(\phi, t) C(t)}{C(\phi)} - 1 \right)^2 . \tag{3.5}$$

3.3.1 Bias parameter estimates.

Among all the bias parameters, the sequence-specific bias is learned in advance while the start and end bias is a function of transcript fragment length. The only bias parameters unknown related to the 3' bias are defined by the intercept γ_0^t and slope γ_1^t for every transcript $t \in \mathcal{T}_g$. Therefore, we use an iterative-minimization strategy and search for a set of bias coefficients γ_0^t's and γ_1^t's that better fit the RNA-seq sample than the uniform sampling model. We start with the transcript coverage $C(t)$'s that are solved from the uniform sampling model (with $\gamma_0^t = 1$ and $\gamma_1^t = 0$ as initial condition). Analogous to the hill climbing algorithm [Russell and Norvig, 2003], we then iteratively probe a locally optimal set of transcript coverage together with the bias coefficients around the uniform solution through minimizing the SSRE. In each iteration, a candidate solution is obtained through sequentially setting the partial derivatives to 0 with respect to every unknown parameter γ_0^t, γ_1^t, $C(t)$, and for every transcript $t \in \mathcal{T}_g$. If the candidate solution results in a smaller SSRE, the candidate solution is taken and the iteration continues.

We use an iterative-minimization strategy to search for a set of bias coefficients γ_0^t's and γ_1^t's for every transcript $t \in \mathcal{T}_g$ that better fit the RNA-seq sample than the uniform sampling model. We initiate the iterations with the transcript coverage $C(t)$'s solved from the uniform sampling model and the bias coefficients $\gamma_0^t = 1$ and $\gamma_1^t = 0$. In each iteration, for transcript t we set:

1. $\frac{\partial SSRE}{\partial C(t)} = 0$; 2.$\frac{\partial SSRE}{\partial \gamma_1^t} = 0$; 3. $\frac{\partial SSRE}{\partial \gamma_0^t} = 0$.

$$
\begin{aligned}
& \frac{\partial SSRE}{\partial C(t)} = 0 \\
\Rightarrow \quad & \sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \sigma(\phi, t)\mathbf{M}'(\phi, t) = 0 \\
\Rightarrow \quad & \sum_{s \in \mathcal{T}_g} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\sigma(\phi, t)\mathbf{M}'(\phi, t)) = \sum_{\phi \in \Phi_g} C(\phi)\sigma(\phi, t)\mathbf{M}'(\phi, t) \\
\Rightarrow \quad & C(t) = \frac{\sum_{\phi \in \Phi_g} C(\phi)\sigma(\phi, t)\mathbf{M}'(\phi, t)}{\sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\sigma(\phi, t)\mathbf{M}'(\phi, t)} \\
& - \frac{\sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\sigma(\phi, t)\mathbf{M}'(\phi, t))}{\sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\sigma(\phi, t)\mathbf{M}'(\phi, t)}.
\end{aligned}
$$

$\sigma(\phi, t)$ is the only function related to γ_1^t and γ_0^t.

$$
\begin{aligned}
& \frac{\partial SSRE}{\partial \gamma_1^t} = 0 \\
\Rightarrow \quad & \sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t)C(t) = 0 \\
\Rightarrow \quad & C(t) \sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t) \\
& = \sum_{\phi \in \Phi_g} C(\phi)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t) - \sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t)).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \frac{\partial SSRE}{\partial \gamma_0^t} = 0 \\
\Rightarrow \quad & \sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t)C(t) = 0 \\
\Rightarrow \quad & C(t) \sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t) \\
& = \sum_{\phi \in \Phi_g} C(\phi)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t) - \sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t)).
\end{aligned}
$$

Because $\sigma(\phi, t)$ is a linear combination of γ_1^t and γ_0^t, and hence $\sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'\phi, t$ is also the linear combination of γ_1^t and γ_0^t. Then we can directly calculate $\frac{\partial \sigma(\phi,t)}{\partial \gamma_1^t}$ and $\frac{\partial \sigma(\phi,t)}{\partial \gamma_0^t}$.

3.3.2 Solving the linear model with LASSO regularization.

Lastly, we solve for the level of individual transcript expression with additional regularization, based on the bias coefficients from the previous step. One common problem in transcript quantification is that the set of expressed transcripts are not known a priori. Hence it becomes crucially important to identify the set of truly expressed transcripts provided in a candidate set. Therefore, we further apply the L1 regularization (known as LASSO) for its proven effectiveness in irrelevance-removal and solve for the set of transcript expression $C(\mathcal{T}_g)$ that minimizes the following loss function

$$L = \text{SSRE} \ + \ \text{L1 penalty} = \sum_{\phi \in \Phi_g} \left(\frac{\sum_{t \in \mathcal{T}_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)C(t)}{C(\phi)} - 1 \right)^2 + \lambda ||C(\mathcal{T}_g)||_1$$

where $\lambda \geq 0$ denotes the weight of the L1 shrinkage and $C(t) \geq 0$ for every $t \in \mathcal{T}_g$.

3.4 Experimental Results

To evaluate the performance of the MultiSplice model, we compared it with four other approaches. The *ExonOnly* model, where only exonic segments are used to represent transcript composition as proposed in SLIDE [Li et al., 2011a], was implemented using a linear regression approach with LASSO. The ExonOnly model provided the baseline comparison for MultiSplice. The *Poisson* model, which was originally proposed by [Richard et al., 2010], was implemented in C. Two read-centric models:

Cufflinks [Trapnell et al., 2010c] which uses the reads aligned to the reference genome and RSEM [Li and Dewey, 2011] which uses the reads aligned to the set of reference transcript sequences are analyzed. Cufflinks 1.1.0 was downloaded from its website in September, 2011. RSEM 1.1.13 was downloaded in November, 2011.

These algorithms were run on both simulated datasets and real datasets. Reads were first mapped by MapSplice 1.15.1 [Wang et al., 2010a] to the reference genome. If the read was paired-end, MapPER [Hu et al., 2010] was applied to infer the alignment of the entire transcript fragment.

3.4.1 Transcriptome identifiability with increasing read length

We first study how the increase in read length may alleviate the lack of identifiability issues in transcript quantification using MultiSplice. We downloaded UCSC gene models in human (track UCSC genes:GRCh37/hg19), mouse (track UCSC Genes: NCBI37/mm9), worm (track WormBase Genes: WS190/ce6) and fly (track FlyBase Genes: BDGP R5/dm3). We computed the feature matrix used in MultiSplice given variable read length and determined its rank. The transcript isoforms of a gene is identifiable if the rank of the feature matrix is no less than the number of transcripts. Figure 3.3 plots the additional number of genes that become identifiable as the read length increases from 50bp assuming single-end read RNA-seq data. For all four species, as the read length increases, MultiSplice is capable of resolving the transcript quantification issues of more genes. With 500bp reads, about 98% genes in both human and mouse become identifiable. Surprisingly, for worm and fly, 500bp reads do not gain significant improvement over 50bp reads. This is mostly due to the

fact that the exon lengths of fly and worm are comparably much longer [Fox-Walsh et al., 2005] than human and mouse, making it difficult for reads of moderate size to take effect. With current short read technology where read length is typically 100bp or less, paired-end reads with the size of transcript fragments around 500bp may be the most economical and effective for transcription quantification for genes with identifiability issues. This is under the assumption that it is possible to infer the transcript fragment from paired-end reads based on the tightly controlled distribution of insert-size.

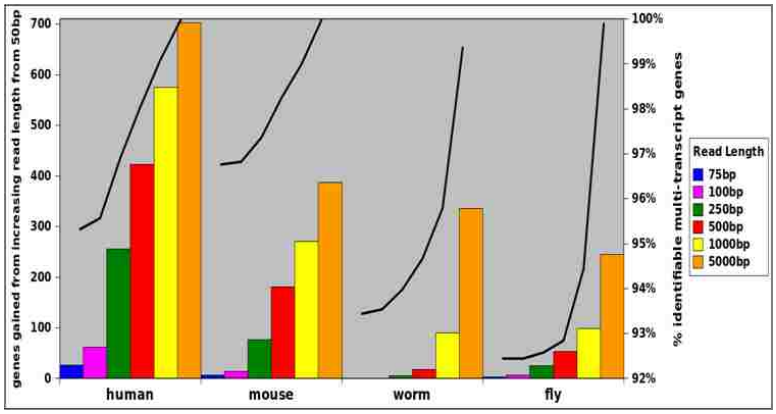


Figure 3.3: Changes in mRNA identifiability as a function of transcript fragment/read length. Starting from levels achieved with 50bp single-end reads, the left side of the y-axis shows the additional number of genes that become identifiable using MultiSplice as the read length increases. The y-axis on the right side shows the total percentage of genes for which mRNA transcript structures are resolved. The UCSC annotated transcript sets of four species: human, mouse, fly and worm were used for this analysis.

3.4.2 Simulated human RNA-seq experiment

Data Simulation. Due to the lack of the ground truth within real datasets, simulated data has become an important resource for the evaluation of transcript quantification algorithms [Bohnert and R., 2010, Li et al., 2010a, Nicolae et al., 2011]. We

developed an in-house simulator to generate RNA-seq datasets of a given sampling depth using UCSC human hg19 annotation. The simulation process consists of three steps: (1) randomly assign relative proportions to all the transcripts within a gene and set this as the true profile; (2) calculate the number of reads to be sampled from each transcript; (3) sample transcript fragments of a given length along the transcripts according to the per base coefficient $\sigma(t,i) = \frac{ki\alpha(t,i)\beta(t,i)}{l(t)} + 1$ for the ith base on transcript t, where $\alpha(t,i)$ and $\beta(t,i)$ are the sequence-specific bias and the transcript start/end bias as defined in Section 4 and k is the slope of the position-specific bias. Paired-end reads will be generating by taking the two ends of the transcript fragment. Please note the sequence bias per base has been learned from a real dataset, a technical replicate of MCF-7 data that will be introduced in the next section.

Accuracy measurement. Due to inconsistencies in the normalization scheme used by different software, the estimated abundance may not be comparable among different approaches. Hence, we computed relative proportions of transcript isoforms for each method. The similarity between the estimated result and the ground truth is measured by both Pearson correlation and Euclidean distance. Pearson correlation is the accuracy measurement used in rQuant [Bohnert and R., 2010]. Let X denote the vector of real isoform proportions of a gene and $\hat{X}$ denote the estimated proportions. The formula of the correlation is: $r(X,\hat{X}) = cov(X,\hat{X})/(\sigma_X \cdot \sigma_{\hat{X}})$. A value close to 1 means that our estimation is highly accurate and vice versa. Below, we adopt a boxplot to illustrate the performance of each method. The box is constructed by the 1st quartile, the median, and the 3rd quartile. The ends of the upper and lower whisker are given by the 3rd quartile $+1.5 \times IQR$*(inner quartile range)* and 1st

quartile $-1.5 \times IQR$, respectively. Due to the space limit, we present the result of correlation measurement in the main manuscript.

Varying read lengths. On the premise of the same sequencing depth, we would like to find out whether or not the read length will affect the estimation results. 40 million RNA-seq fragments were simulated from the human transcriptome. 2x50bp paired-end reads (insert size around 150bp) were generated from these fragments. A 50bp single-end read set was constructed by simply throwing out the second read of each pair and the 100bp single reads were obtained by taking the 100bp prefix of the transcript fragments. This configuration allows a fair evaluation about the effect of varying read lengths by eliminating difference from random read sampling.

As shown in Figure 3.4, the performance of MultiSplice, RSEM and ExonOnly method improves as the read length increases. Accuracy of the Poisson model does not change much with varying read lengths. It is surprising to see that Cufflinks achieves better correlation with 100bp single-end reads than both 2x50bp paired-end reads and 50bp single-end reads. This is probably because the transcript fragment inference from paired-end reads may not be accurate for Cufflinks. Both MultiSplice and RSEM show higher median correlation and lower variance compared with other methods under different read lengths, which indicates that MultiSplice and RSEM are capable of leveraging longer reads for more accurate estimation as RNA-seq technologies improve.

Varying sampling depth. Next we evaluate how the sequencing depth may affect the accuracy of transcript abundance estimation. Four groups of 2x50bp paired-end synthetic data were generated on the whole human transcriptome with increasing

number of reads: 5 million, 10 million, 20 million and 40 million. Since the exonic regions of different genes may overlap, we quantify isoforms within a genomic locus [Trapnell et al., 2010c]. 13364 genomic loci with multiple isoforms are selected for analysis. The loci were divided into three subsets: (1) 12413 loci to which identifiability holds for all methods; (2) 498 loci to which identifiability holds for MultiSplice; (3) 453 loci to which identifiability does not hold for all methods.

For each subplot in Figure 3.5(a, b, c), the estimation accuracy for all methods generally improves as more reads are sampled. For the loci whose identifiability conditions are satisfied for all methods, the estimated transcript proportion is highly similar with the ground truth, with an median correlation close to 0.9 for all methods. In the second category, when the genes are still identifiable with MultiSplice, the estimation accuracy of MultiSplice and RSEM remain high, with an median correlation above 0.8 while others slip below 0.7. For the category when identifiability is not satisfied for all methods, the estimation accuracy is degraded even more. However,

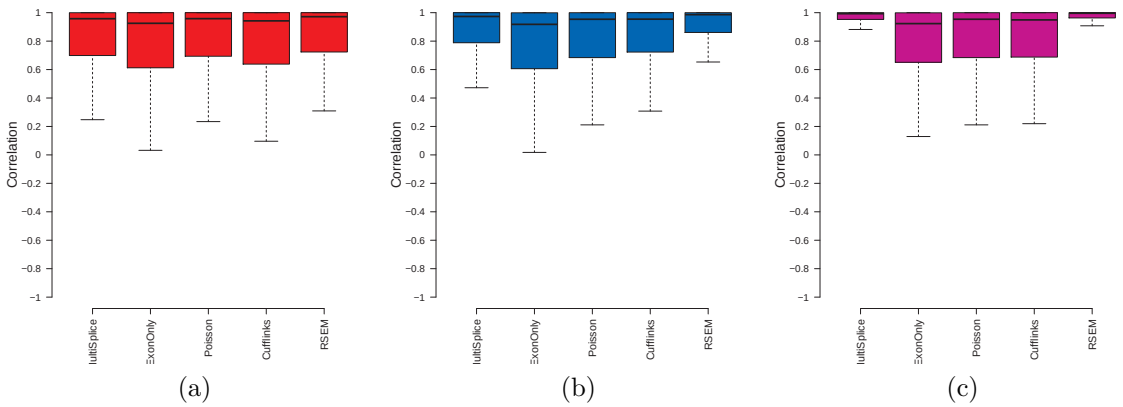


Figure 3.4: a-c. Boxplots of the correlation between estimated transcript proportions and the ground truth under varying read length. (a),(b) and (c) correspond to the estimation results on 40M 50bp single-end reads, 40M 100bp single-end reads, and 40M 2x50bp paired-end reads, respectively.

MultiSplice still consistently gives better estimation results indicating that the inclusion of MultiSplice features make transcript quantification more stable than other methods. Cufflinks demonstrated the worst performance in this category with largest variance as also shown in Figure 3.7(c), mainly because the unidentifiability conditions make it difficult to assign these reads to a transcript. Instead, it throws out

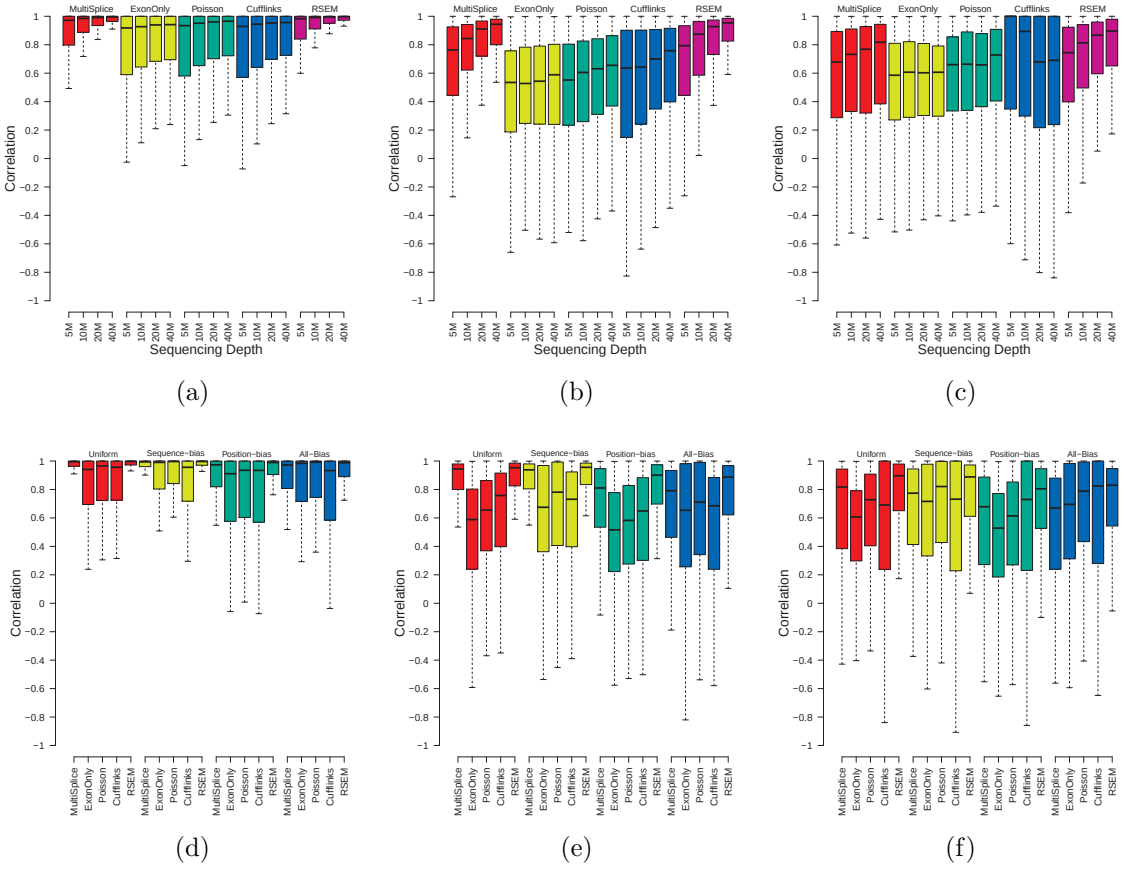
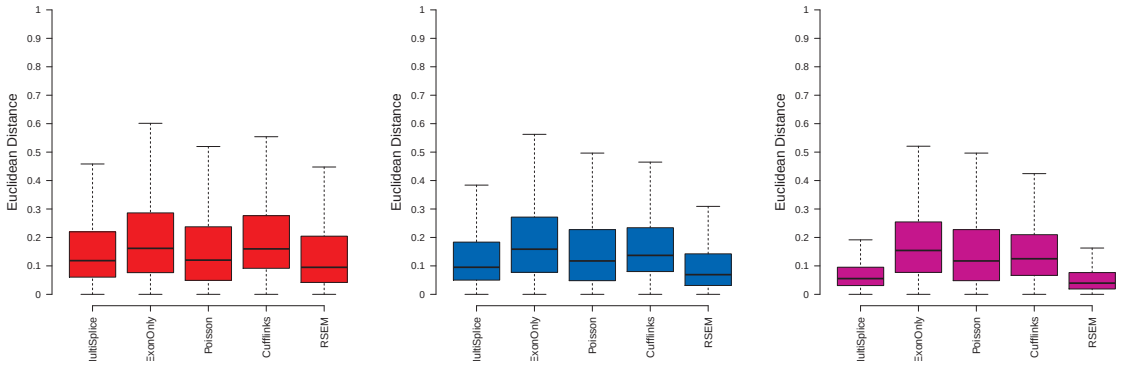


Figure 3.5: a-c. Boxplots of the correlation between estimated transcript proportions and the ground truth under varying number of sampled reads: 5M, 10M, 20M and 40M over a total of 13364 genomic loci with more than one isoforms. (a), (b) and (c) correspond to the loci set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively. d-f: Boxplots of the correlation between estimated transcript proportions and the ground truth under four circumstances: uniform sampling, sampling with positional bias only, with sequence bias only and with all bias. (d), (e) and (f) correspond to the loci set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.



(a) (b) (c)

Figure 3.6: a-c. Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth under varying read length. (a), (b) and (c) correspond to the estimation results on 40M 50bp single-end reads, 40M 100bp single-end reads, and 40M 2x50bp paired-end reads, respectively.

most of multi-mapped reads. Apparently, increasing sampling depth cannot alleviate the issue of unidentifiability.

Bias correction. To study the effect of the bias correction, we have simulated data with uniform sampling, sampling with only positional bias, sampling with only sequence bias, and sampling with the combined positional and sequence bias. Here, we set the slope of the position-specific bias k to 2 with 40 million 2x50bp paired-end reads sampled from the whole transcriptome for each case. All the approaches achieve the best results when the sampling process is uniform. As positional or sequence bias is introduced, their performance tapers down. The presence of both positional and sequence biases has the largest impact in all methods. Meanwhile, because MultiSplice and Cufflinks correct both sequence and positional bias and RSEM could adjust positional bias, these three methods are more robust and outperform the ExonOnly and the Poisson methods.

Inference of expressed transcripts. Quantification of mRNAs usually relies

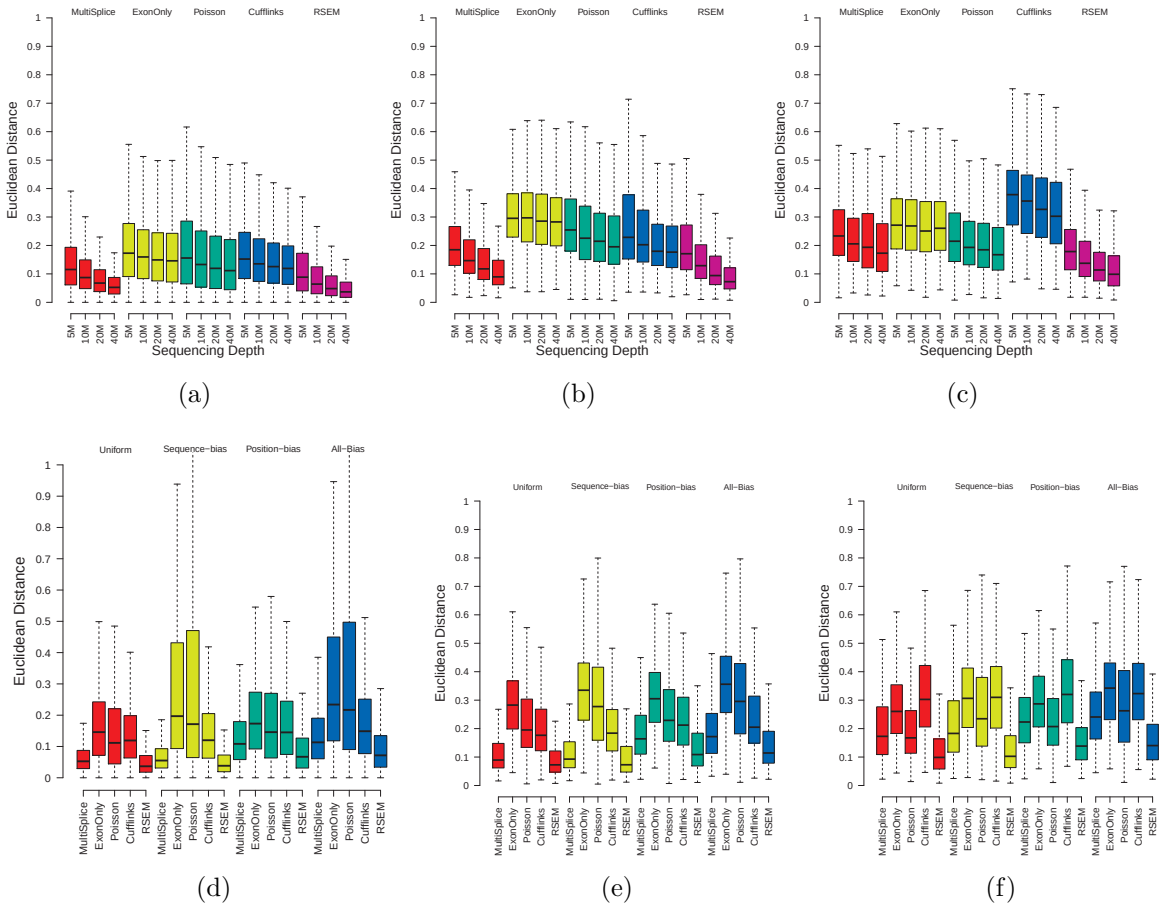


Figure 3.7: a-c. Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth under varying number of sampled reads: 5M, 10M, 20M and 40M over a total of 13364 genomic loci with more than one isoforms. (a), (b) and (c) correspond to the loci set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively. d-f: Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth under four circumstances: uniform sampling, sampling with positional bias only, with sequence bias only and with all bias. (d), (e) and (f) correspond to the loci set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.

on a set of candidate transcript structures as input. It is unknown in a priori whether each transcript is present in a sample or not. Therefore, accurate quantification methods should be able to infer the transcripts that are expressed as well as those that are not. To assess the capability of the various methods to infer expressed transcripts,

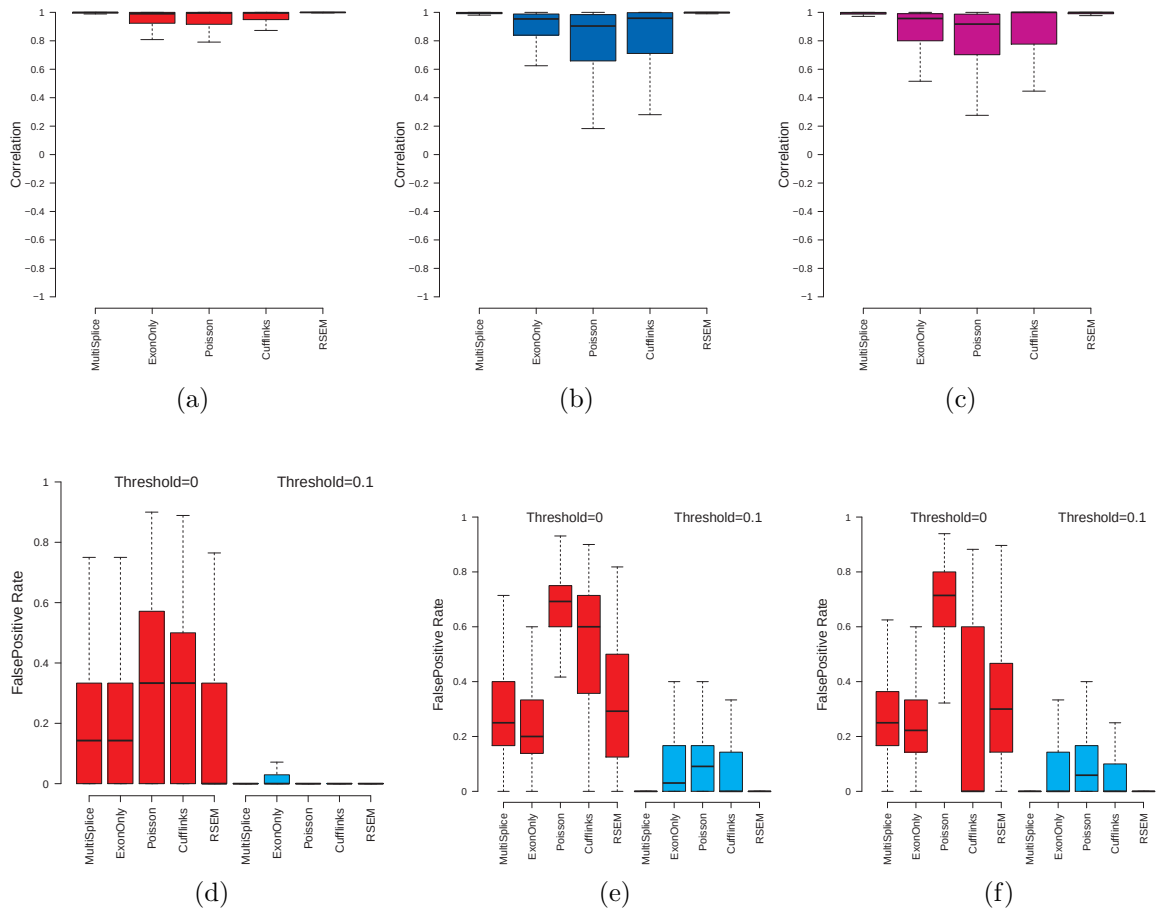
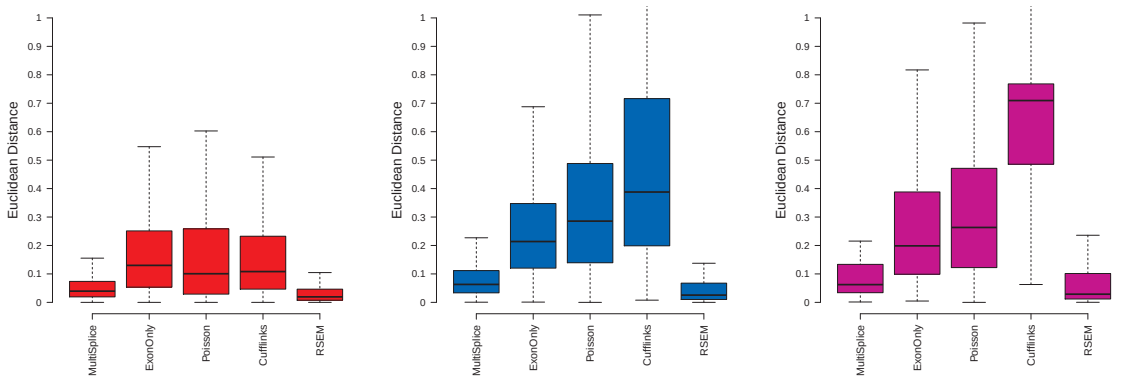


Figure 3.8: a-c. Boxplots of the correlation between estimated transcript proportions and the ground truth. (a), (b) and (c) correspond to the loci set that is identifiable with the basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively. d-f. Comparison of false positive rates in the inference of the expressed transcripts. Thresholds represent the minimum fraction of a transcript that is considered expressed. (d), (e) and (f) correspond to the loci set that is identifiable with the basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.



(a) (b) (c)

Figure 3.9: a-c. Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth for inference of dominant transcripts. (a), (b) and (c) correspond to the loci set that is identifiable with the basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.

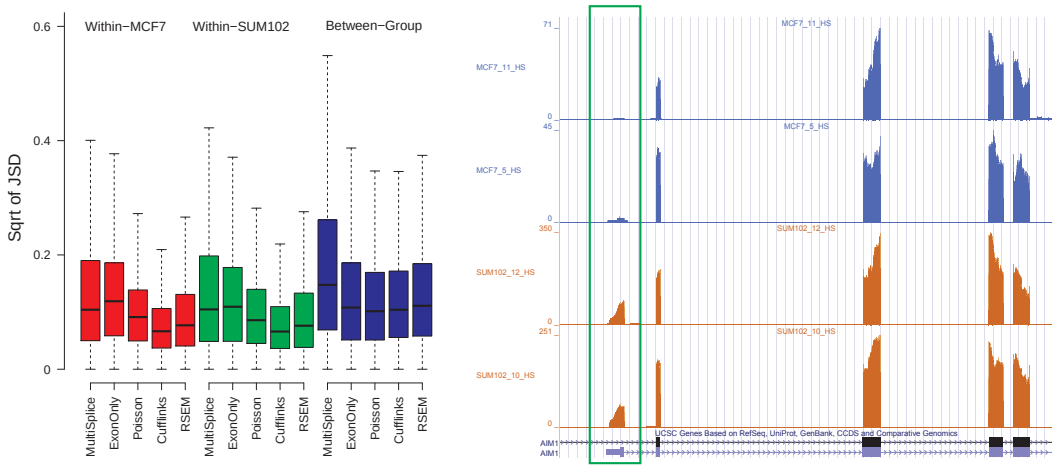
we generated 40 million simulated 2x50bp paired-end reads from human genes with at least 3 transcripts. We randomly chose two transcripts from one gene and simulated reads only from these transcripts. The remaining transcripts were not sampled. We used the false positive rate to measure the accuracy of the inference. Non-expressed transcripts that were estimated with a positive abundance above a given threshold were counted as the false positives. As shown in Figure 3.8(a, b, c), MultiSplice and RSEM demonstrated best estimation accuracy and further more MultiSplice demonstrated the lower false positive rate in the identification of dominant transcripts in Figure 3.8(d, e, f). Poisson and Cufflinks tended to assign positive expression to every transcript including those that are not expressed. MultiSplice, in general, outperformed the others in identifying the correct set of expressed transcripts.

3.4.3 Real human RNA-seq experiment

We applied the set of transcript quantification methods to a dataset that was originally used by Singh et al. to study differential transcription [Singh et al., 2011]. In this study, two groups of RNA-seq datasets were generated from SUM-102 and MCF-7, two breast cancer cell lines. Each group contains 4 samples as technical replicates. The RNA-seq data were generated from Illumina HISEQ2000. Each sample had approximate 80 million 100bp single-end reads. About 60 million reads can be aligned to the reference genome by MapSplice. The UCSC human hg19 annotated transcripts were fed into each software for transcript quantification.

Since ground truth expression profiles do not exist for the real datasets, we investigated whether the different methods provided a consistent estimation within samples of technical replicates which only vary by random sampling. In contrast, a significant number of genes between MCF-7 and SUM-102 were expected to be differentially expressed [Singh et al., 2011]. To evaluate this, we computed Jensen−Shannon divergence (JSD), used in Cuffdiff [Trapnell et al., 2010c] to measure the dissimilarity between two samples and calculated the *within-group* and *between-group* differences. As detailed in Figure 3.10(a), MultiSplice, Cufflinks and RSEM had smaller average within-group difference than the average between-group difference while the other two methods do not show clear difference. MultiSplice demonstrated higher between-group difference than both Cufflinks and RSEM, but also had relatively higher within-group differences as well. Most of these, however, were well below a JSD of 0.2 and considered to be insignificant. A closer look at a number of cases

showed that occasionally MultiSplice and Cufflinks may overestimate or underestimate the between-group difference respectively. Figure 3.10(b) (The complete figure with 8 samples can be found in the Figure 3.11(a)) shows a gene where Cufflinks underestimated the difference between the two groups. The second isoform of the gene AIM1 has a unique first exon (chr6:106989461-106989496). Clear difference in the read coverage on this exon can be observed between the two groups, indicating strong differential levels of expression, i.e., the second isoform is barely expressed in MCF-7 while almost comparable to the first isoform in SUM-102 cells. The between group square root of JSD is 0.21 by Cufflinks, lower than 0.39 by RSEM and much lower than 0.50 by MultiSplice.



(a) (b)

Figure 3.10: **a.** Boxplots of the within-MCF-7, within-SUM-102, and between-group square root of JSD of all genes for all methods. **b.** A case where Cufflinks underestimated the difference between the two groups. The second isoform of Gene AIM1 has a unique first exon, whose read coverage differs significantly between the two groups. A detailed plot with all 8 samples can be found in the Figure 3.11(a).

The exon-skipping event found in gene CD46 is also differentially expressed (Figure 3.11(b)). The estimation of transcript quantification with MultiSplice was con-

sistent with the observation in the qRT-PCR data showing that steady state levels of transcripts with the skipped exon were present in amounts more than two fold higher expression in SUM-102 than in MCF-7 cells. An additional example is shown in Figure 3.12.

Computational Performance. We also compared the running time and memory usage of the proposed method with Cufflinks and RSEM. In order to make a fair comparison, we only measured the computational performance of transcript quantification for each software. One sample with 76 million reads from MCF-7 was used for analysis. The reads are aligned to the reference transcript set by Bowtie [Langmead et al., 2009a] for RSEM and to the reference genome by MapSplice [Wang et al., 2010a] for MultiSplice and Cufflinks. The results presented here were run on Intel Xeon X5650 (Westmere) 12-core 2.66 GHz Linux server with 32GB of RAM and single-thread enabled.

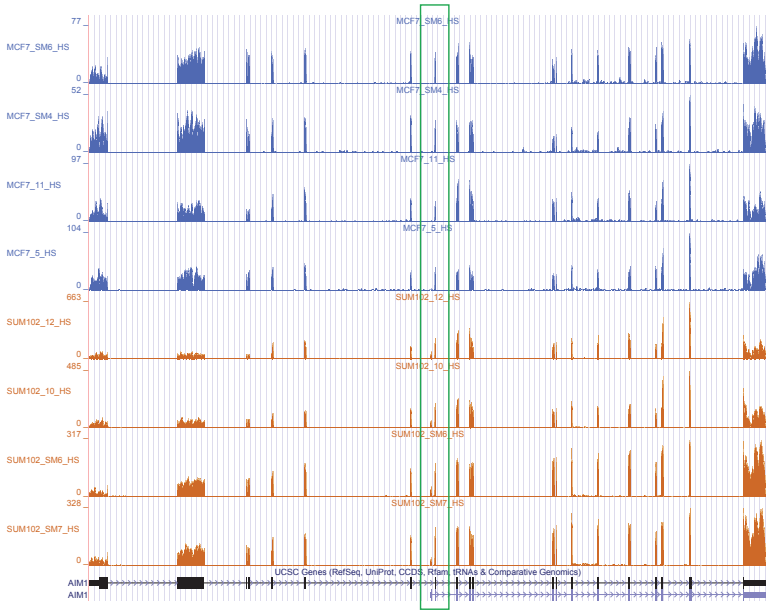
Table 3.1 summarize the comparison results of MultiSplice, Cufflinks and RSEM.

Table 3.1: Computational performance comparison

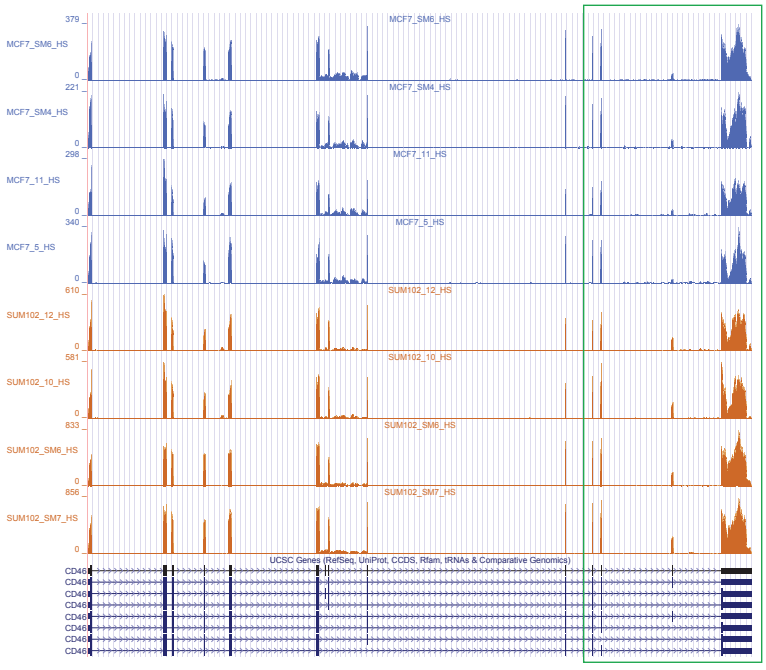
Method Type	MultiSplice	Cufflinks	RSEM
QuantificationTime	$40min$	$74min$	$23h$
MemoryUsage	$< 1G$	$2G$	$7G$

3.5 Discussion

In this chapter, we have presented a general linear framework for the accurate quantification of alternative transcript isoforms with RNA-seq data. We introduce a set of new structural features, namely MultiSplice, to ameliorate the issue of *identifiability*.



(a)



(b)

Figure 3.11: a. The coverage plot of Gene AIM1 in all 8 breast cancer cell line samples. Please note the first exon of the second isoform is barely expressed MCF-7 but its expression significantly increased in the SUM-102 samples. b. The coverage plot of Gene CD46. The exon-skipping event on the 13th exon has been confirmed by qRT-PCR.

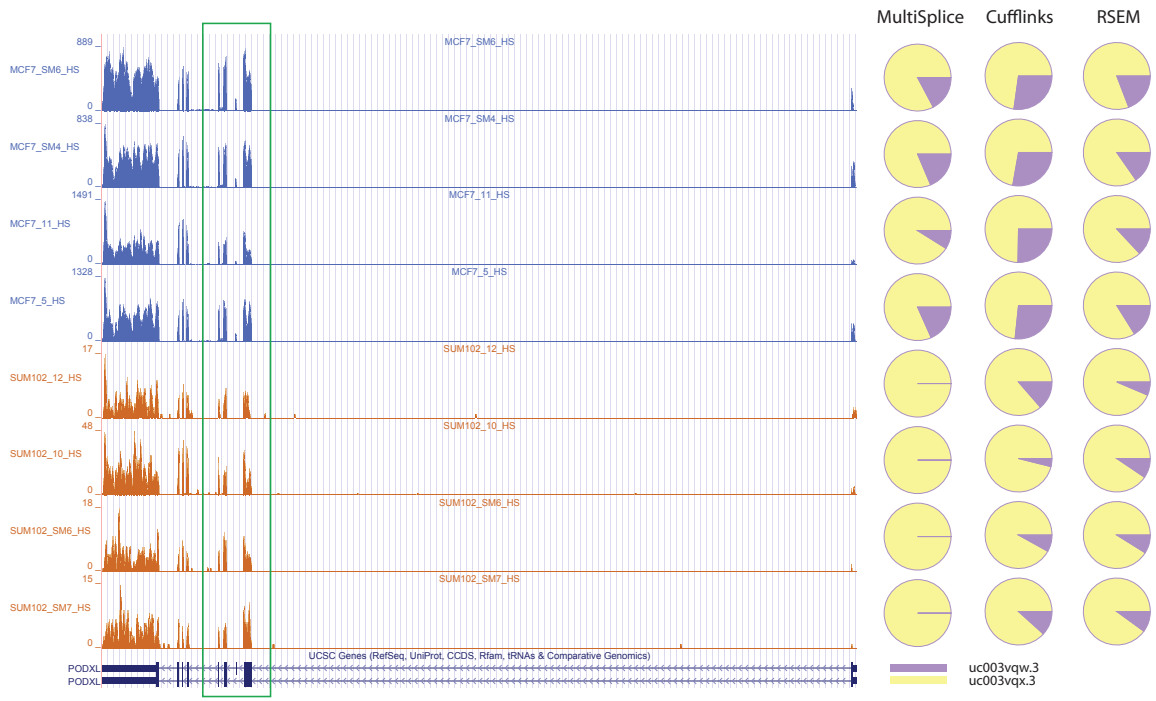


Figure 3.12: One real gene example for which MultiSplice inferred the expressed transcript while RSEM and Cufflinks failed to do so. The left figure shows the coverage plot of Gene PODXL in all 8 breast cancer cell line samples. The between group square root of JSD is 0.290611 by MultiSplice, 0.195271 by Cufflinks and 0.094207 by RSEM. The exon-skipping event on the seventh (chr7: 131194995-131195090) are differentially expressed between two cell lines. The coverage plot indicates the first isoform is not expressed in SUM-102. The right part shows pie charts of estimated relative expression of the annotated two isoforms for three methods in all 8 samples. Except MultiSplice, both Cufflinks and RSEM assign positive expression to the first isoform in SUM-102.

With MultiSplice features, 98% of UCSC gene transcript models in human and mouse become identifiable with 500bp reads (or paired-end reads with 500bp transcript fragments), an 8% increase from 50bp. Therefore, longer reads or paired-end reads with longer insert-sizes rather than further increases in sequencing depths can be crucial for the accurate quantification of mRNA isoforms with complex alternative transcription, even though a majority of the genes have relatively simple transcript variants. The results also demonstrate the robustness of the MultiSplice method under various

sampling biases, consistently outperforming three other methods: Cufflinks, Poisson and ExonOnly and comparable to RSEM. The application of our approach to real RNA-seq datasets for transcriptional profiling successfully identified a number of isoforms whose proportion changes differed significantly between two distinct breast cancer cell lines.

Chapter 4 Simultaneous Transcript Reconstruction and Quantification

In this chapter, we revisit the problem of transcript reconstruction. Instead of assigning reads probabilistically to a set of isoforms, we go one step further by answering how well individual reads may be pieced together to build copies of individual transcripts. We directly reconstruct *effective* transcript copies, each of which corresponds to a chain of non-overlapping transcript fragments (Figure 4.1). The contribution of each effective copy to the abundance of the corresponding isoform does not solely depend on the number of reads observed, but also on how consistent the distribution of the observed fragments is as compared to the expected process of mRNA fragmentation and sampling. This procedure allows us to explicitly take into account of the positional relationships among reads, which were generally ignored by existing methods. In the meantime, the total number of transcript copies constructed can be used to assess the transcript abundance. Therefore, we introduce a new measure for transcript quantification, namely *eTPM*, **e**ffective Transcripts Per Million. The eTPM of an isoform i is calculated as:

$$\mathrm{eTPM}_i = \frac{\mathrm{eT}_i \times 10^6}{\sum_j \mathrm{eT}_j} \tag{4.1}$$

where eT_i is the number of effective copies of isoform i and $\sum_j \mathrm{eT}_j$ accounts for the total number of transcripts in the transcriptome. With our approach, not only do the constructed effective copies convey the information of the exon composition of the transcript, but the number of copies also delivers an estimation of the relative

abundance of each isoform. It is therefore truly *simultaneous* in terms of transcript identification and quantification.

To this end, we have developed a novel computational algorithm *Astroid*(Transcript reconstruction through assembly of effective transcript copies guided by the fragment distance.). We model the relation of all observed reads using a directed flow network, with reads connected by edges whose weight represents the likelihood that two reads may coexist in a transcript. The most likely set of transcript copies is reached by solving a min-cost flow problem given the flow network. A compression scheme is developed to speed up the performance for genes with high read coverage. The model is further consolidated by adding MultiSplice features [Huang et al., 2012], reads that span multiple alternative splicing events, to avoid the identification of spurious transcripts.

We have compared the performance of our method with a number of state-of-the-art methods including Cufflinks [Trapnell et al., 2010b], Scripture [Guttman et al., 2010], IsoLasso [Li et al., 2011b] and Trinity [Grabherr et al., 2011]. Simulation studies on the human transcriptome datasets have demonstrated Astroid's superior sensitivity and precision on transcript discovery. The eTPM estimate calculated from the number of effective transcript copies assembled by Astroid has exhibited an improved correlation with true transcript abundance than FPKM estimates. The evaluations on the MAQC human brain dataset and the Alexa-seq dataset further demonstrated the effectiveness of our method in real applications, in which Astroid provided slightly more consistent estimates for transcript abundance with qRT-PCR validations than other methods.

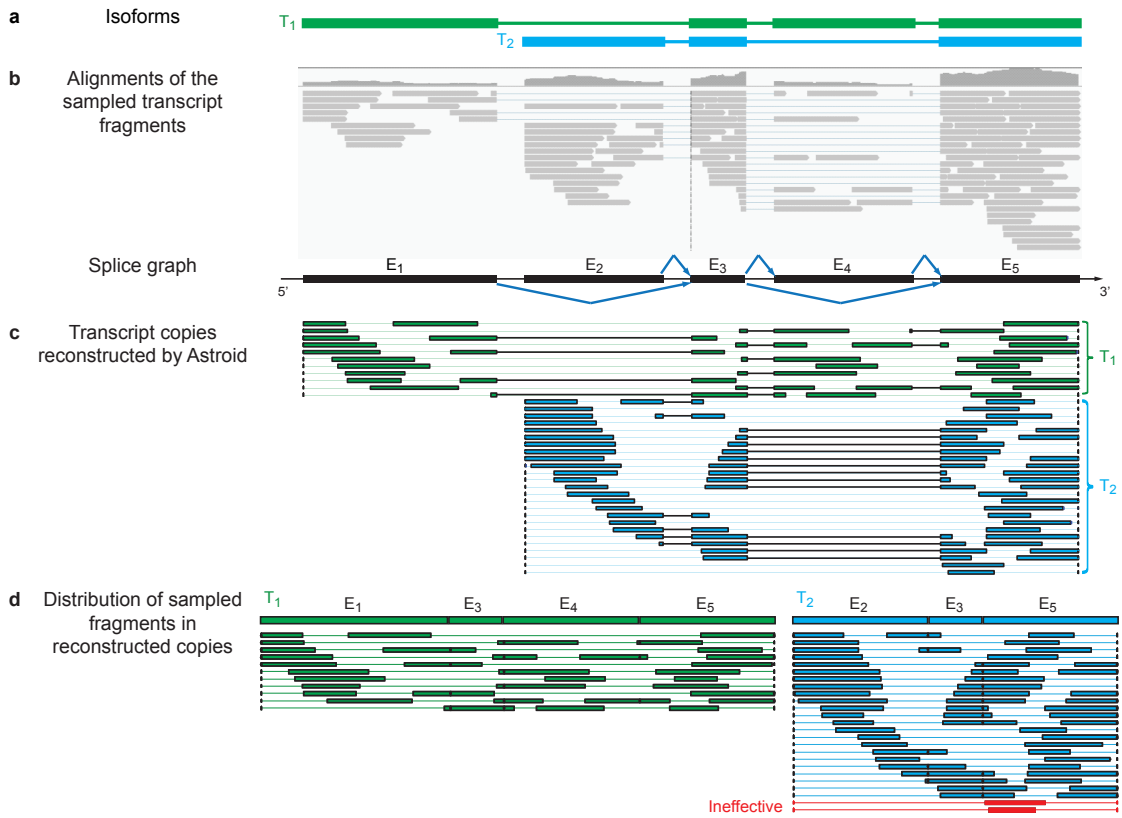


Figure 4.1: Reconstruction of effective transcript copies by Astroid. **(a)** The two isoforms from which transcript fragments are randomly sampled. **(b)** The alignments of the sampled fragments, plotted with IGV Thorvaldsdóttir et al. [2013]. A splice graph can be built based upon the exons and splice junctions identified from the fragment alignments. **(c)** Effective transcript copies assembled by Astroid. Astroid successfully reconstructs the two expressed isoforms with no false positive. **(d)** The distribution of fragments in the effective copies. The likelihood of each copy is assessed according to the sizes of the fragments in the copy together with the between-fragment distances. Effective transcript copies will be identified and used to measure the abundance of each isoform. Note that this example shows only transcript fragments rather than the RNA-seq reads for simplified illustration. However, our method does take paired-end reads as input.

4.1 Effective Transcripts

We propose to construct a set of *effective transcript copies* which simultaneously explain the observed reads and estimate the transcript abundance. The notations used in this section and the following method sections are summarized in Table 4.1.

Table 4.1: Notations in the main manuscript.

Symbol	Meaning
i	isoform
t	transcript copy
N_i	read count on isoform i
N	total read count in the transcriptome
$len(\cdot)$	length of $\cdot$ (read, transcript, etc.)
R_t^{fr}	set of sampled fragments in copy t
R_t^{gap}	set of between-fragment gap in copy t
$d(\cdot)$	fragment/gap size distribution
eT_i	effective transcripts of isoform i
$P_W(\cdot\|\delta_i, \eta)$	characterized Weibull distribution
$\phi(v)$	exon is splice graph a read v is mapped to
$\rho(\phi(v_1), \phi(v_2))$	a splice graph path between two exons
$f(e)$	the amount of flow on edge e
γ	compression parameter
π	read cluster
b	MultiSplice feature
$c(e)$	capacity on edge e in RFN
$\psi(b)$	size of the sampling window of b

The sampled fragments typically do not immediately follow each other and fragments may not be sampled immediately at the start/end of a transcript. We model the positional relationship among the fragments by considering the *size distribution* of 1) the fragments, 2) the gap between two adjacent fragments, and 3) the gap from the transcription start site to the first fragment and from the last fragment to the transcription termination site. We use R_t^{fr} to denote the set of transcript fragments in a copy t. Each fragment can be identified by a mate pair of reads. The set of between-fragment gaps in t is denoted as R_t^{gap}.

The likelihood of the transcript t is then interpreted as the joint likelihood of all its fragments together with the gaps according to their sizes. To simplify the model, we assume that the sizes of the fragments and the gaps follow the same distribution,

whose density function is denoted as $d(\cdot)$.

Let isoform i be the isoform that t is copied from, denoted as $t \in i$. Because different copies may have different number of component fragments, we use the geometric mean of the probabilities of all fragments and gaps in t to evaluate the likelihood of t,

$$\mathcal{L}(t) = \left(\prod_{r \in R_t^{gap}} d(len(r)) \prod_{r \in R_t^{fr}} d(len(r)) \right)^{\frac{1}{|R_t^{gap}|+|R_t^{fr}|}}.$$

Generally, $\mathcal{L}(t)$ represents a central tendency of the probability of the fragments and gaps contained in t. It is possible to model the size distribution of the gaps differently, with more complex distributions. However, as the experimental results have suggested, the approximation in our simplified model is sufficient. The distribution of $d(\cdot)$ will be discussed later in this section.

We further determine the effectiveness of a copy t by assessing the probability of observing a copy with likelihood no greater than $\mathcal{L}(t)$.

Definition For a transcript t, let $\mathcal{I}_t$ denote the subdomain of the size density $d(\cdot)$ such that $\forall x \in \mathcal{I}_t$, $d(x) \leq \mathcal{L}(t)$. Then t is **effective** if the cumulative density integrated over all $x \in \mathcal{I}_t$ is no greater than a significance level τ, i.e., $\int_{\mathcal{I}_t} d(x)\, dx \leq \tau$, where τ controls the probability of falsely considering t as ineffective.

As a convention in hypothesis testing, τ is often set as 0.05. For a given τ, the set of all effective transcript copies of isoform i is denoted as: $S_i = \{t : \int_{\mathcal{I}_t} d(x)\, dx \leq \tau\}$. Then the abundance of isoform i is measured by the number of its effective copies, called *effective transcripts* (eT), $\mathrm{eT}_i = |S_i|$.

Under the assumption of a uniformly random fragmentation process, the size distribution of the fragments generated from isoform i can be approximated as a characterized Weibull distribution [Griebel et al., 2012, Tenchov et al., 1985] with two parameters δ_i and η. The isoform-specific shape parameter δ_i depends on the logarithm of the molecule length of i, and the scale parameter η reflects the fragmentation intensity which is constant across all transcripts in one experiment [Griebel et al., 2012]. In this paper, the distribution of $d(\cdot)$ is approximated using the Weibull distribution, $d(\cdot) = P_W(\cdot|\delta_i, \eta)$.

4.2 Effective Transcripts per Million (eTPM)

We define the relative expression estimate *effective transcripts per million* of isoform i by normalizing eT_i by the total effective transcript copies in the transcriptome (Equation 4.1).

There exist two other measurements focusing on quantifying the relative isoform expression levels. They are both based on the number of reads on the isoform. One is FPKM [Trapnell et al., 2010b]. For an isoform i, it approximates the transcript abundance by normalizing the number of fragments on the isoform N_i by the isoform length $len(i)$, and uses the total number of fragments per million as a measure of total transcripts in the transcriptome (Equation 4.2). However, when comparing the isoform abundance among samples, the latter approximation is not accurate due to the variant size distribution of the transcripts among different samples [Wagner et al., 2012].

$$\text{FPKM}_i = \frac{N_i}{\frac{len(i)}{10^3} \cdot \frac{N}{10^6}} = \frac{N_i \cdot 10^9}{len(i) \cdot N} \tag{4.2}$$

Another measure TPM [Li and Dewey, 2011, Wagner et al., 2012], *i.e.*, Transcripts Per Million, resolves the inconsistency problem. It approximates the transcript number by normalizing the cumulative per base read coverage by the isoform length. TPM of an isoform i is then calculated as in Equation 4.3 with summing up the estimated abundance of all isoforms accounting for the total number of transcripts in the transcriptome.

$$\text{TPM}_i = \frac{\frac{N_i len(r)}{len(i)} \cdot 10^6}{\sum_j \frac{N_j len(r)}{len(j)}} = \frac{N_i len(r) \cdot 10^6}{len(i) \cdot \sum_j \frac{N_j len(r)}{len(j)}} \tag{4.3}$$

Here $len(r)$ refers to the expected fragment length.

However, it is unclear how well $N_i \times 10^3/len(i)$ in FPKM and $N_i len(r)/len(i)$ in TPM can approximate the true abundance of one isoform because it is impossible that all observed fragments can be tightly arranged one after the other (Figure 4.1b) making every single base of the isoform covered by the read.

Unlike FPKM or TPM, eTPM explicitly considers the possible gaps between two adjacent fragments on the same transcript copy. Since eTPM is normalized by the total number of transcripts in a sample, it can be invariant across samples [Wagner et al., 2012]. While both FPKM and TPM treat each read independently and consider them as the same, the effective transcripts used in eTPM is assessed according to the distribution of their fragments. In real experiments, the position distribution of

sampled fragments may not be uniform due to PCR amplification error [Aird et al., 2011] or sampling biases [Li et al., 2010b, Roberts et al., 2011, Turro et al., 2011]. The affected reads will form ill transcripts copies with only small fractions sampled. These ill transcripts will be recognized during the eTPM calculation, which allows for a more robust abundance measure (Figure 4.1d).

Although this measurement relies on the quantity of assembled effective copies rather than the number of reads, it is derived based on the same assumption as the other measurements regarding the abundance. Longer transcripts require more reads to construct an effective copy. Hence eTPM of different transcripts can be compared directly without the normalization by transcript length.

4.3 Method

The assembly of effective transcript copies with RNA-seq reads is achieved by solving a minimum-cost flow problem. In this section, we detail the modeling of the problem, its solution and various improvements over the basic approach. The input to our method is the genomic alignments of the paired-end reads to the reference genome [Trapnell et al., 2009a, Wang et al., 2010a]. Another important data structure we used is the splice graph [Heber et al., 2002, Hu et al., 2012, Rogers et al., 2012, Xia et al., 2011] (Figure 4.1b). The splice graph is constructed directly from the read alignments using the method described by Hu *et al.* [Hu et al., 2012], and will be used to infer potential transcripts where a pair of reads come from. In general, the exons are identified as the genomic regions covered by abundant reads. These exons constitute the vertices of the splice graph. The spliced read alignments contain splice

junctions, each of which spans a pair of exons. The splice junctions make the edges in the splice graph, whose directions can be defined by the direction of the transcription. In addition, the donor and acceptor sites of a splice junction also determine the boundaries of an exon. A path in the splice graph corresponds to (part of) a possible isoform.

4.4 Read Flow Network

We model the relationships among reads using a flow network, namely the Read Flow Network $RFN = \langle V, E, W, source, sink \rangle$. The vertex set V corresponds to the union of the set of reads and the set of transcription start/termination sites (The transcription start sites and termination sites can be either inferred as the genomic positions that exhibit certain characteristic signatures [Kapranov, 2009, Yamashita et al., 2011] or provided from existing transcript annotation). There are two types of edges between two read vertices, the *in-fragment* edges and the *between-fragment* edges. The in-fragment edges (denoted as E^{in}) refer to edges between reads generated from the same fragment. In the case of paired-end reads, the edge is between the two mates of a paired read. The between-fragment edges (denoted as E^{btwn}) refer to the edges that connect one fragment with its downstream fragment. In this case, an edge usually connects the 3' end read of a fragment (or a transcription start site) to the 5' end of a downstream fragment (or a transcription termination site). Let $\phi(v)$ be the exon in the splice graph where a read v is aligned to. For two vertices $v_1, v_2 \in V$, There exists an edge between v_1 and v_2 for each unique path $\rho_i(\phi(v_1), \phi(v_2))$ between exon $\phi(v_1)$ and exon $\phi(v_2)$ in the splice graph. In presence of alternative splicing,

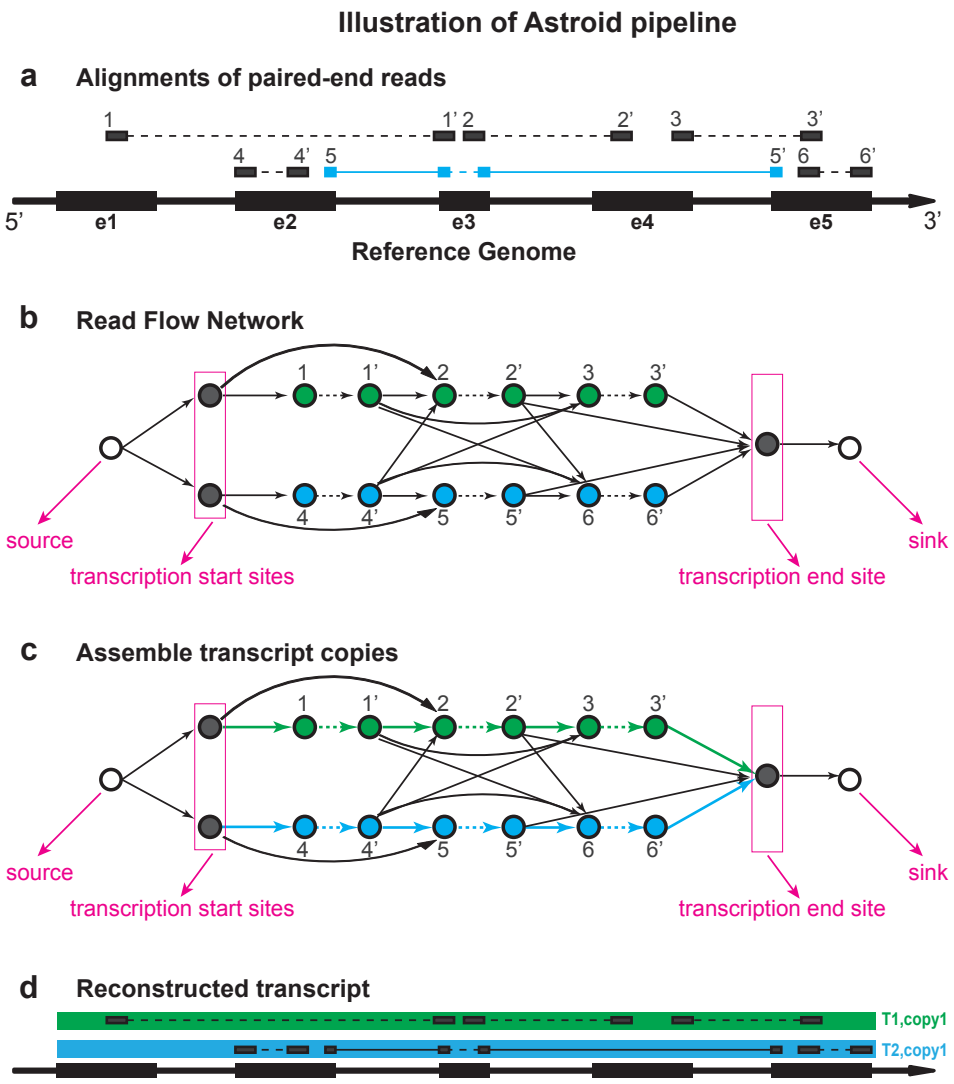


Figure 4.2: **(a)** Alignments of the sequenced paired-end RNA-seq reads on the reference genome. **(b)** The read flow network that relates reads with in-fragment edges (dashed arrows) and between-fragment edges (solid arrows). **(c)** Solve a minimum flow (colored) on the read flow network. **(d)** The assembled effective transcript copies with maximized likelihood.

there may exist more than one paths in the splice graph from $\phi(v_1)$ to $\phi(v_2)$. In this case, multiple edges may be added to include all paths.

The weight of an edge e, $e \in E$ corresponding to a path $\rho_i(\phi(v_1), \phi(v_2))$ between two reads v_1 and v_2, reflects the likelihood of the two reads coming consecutively from the same transcript copy. It is evaluated by the probability of observing a

portion between v_1 and v_2 on path $\rho_i(\phi(v_1), \phi(v_2))$. For $e \in E^{in}$ and $e \in E^{btwn}$, $len(e)$ denotes the size of the observed fragment and size of the between-fragment gap, respectively. Assuming both sizes follow a Weibull distribution, then the probability of e is calculated as $P(e) = P_W(len(e))$. The weight of e in RFN is defined as the negative logarithm of its likelihood, $w(e) = -\log P(e)$.

Lastly, the network RFN is augmented by adding a virtual *source* and a virtual *sink* to initiate and terminate all transcript copies. Directed edges will be built from *source* to all vertices that correspond to transcription start sites and from all vertices that correspond to transcription termination sites to *sink*. Moreover, one edge is added to connect *source* to *sink*. The weights on these edges are always set as 0. Because every read may only originate from one transcript copy, the capacity constraint on every vertex that represents a read is set as 1. The capacities on *source*, *sink* and vertices that represent transcription start/termination sites all equal to the number of vertices that represent the reads.

In this way, each transcript copy can be represented as a *source* to *sink* path (flow) (Figure 4.2b). Let $\mathcal{T}$ denote one set of transcript copies in RFN. For every copy $t \in \mathcal{T}$, the likelihood of t can be evaluated as the product of the probabilities of its reads (vertices included in t), the probabilities of its distances connecting paired-end reads (*in-fragment* edges in t) and the probabilities of its distances connecting transcript fragments (*between-fragment* edges in t). The transcript copies in $\mathcal{T}$ are considered mutually independent because the vertices and edges included in one copy are exclusive. Hence, the likelihood of $\mathcal{T}$ can be written as the joint probability of all the transcript copies in $\mathcal{T}$,

$$P(\mathcal{T}) = \prod_{t\in\mathcal{T}} P(t) \tag{4.4}$$

$$= \prod_{t\in\mathcal{T}} \prod_{v\in t\cap V} P(v) \prod_{e\in t\cap E^{in}} P(e) \prod_{e\in t\cap E^{btwn}} P(e).$$

The probability of a read $P(v)$ can be calculated by considering the quality of its alignment quality [Li and Dewey, 2011]. The probability of an edge $P(e)$ has been defined as $P_W(len(e) = l|e)$. Then the maximum likelihood set of transcript copies can be written as,

$$\hat{\mathcal{T}} = \arg\max_{\mathcal{T}} \log P(\mathcal{T}) \tag{4.5}$$

$$= \arg\max_{\mathcal{T}} \sum_{t\in\mathcal{T}} \left(\sum_{v\in t\cap V} \log P(v) + \sum_{e\in t\cap E^{in}} \log P(e) + \sum_{e\in t\cap E^{btwn}} \log P(e) \right)$$

$$= \arg\min_{\mathcal{T}} \sum_{t\in\mathcal{T}} \left(\sum_{v\in t\cap V} -\log P(v) + \sum_{e\in t\cap E^{in}} -\log P(e) + \sum_{e\in t\cap E^{btwn}} -\log P(e) \right).$$

Therefore, the problem of solving the maximum likelihood set of transcript copies is equivalent to a *minimum-cost flow* problem [Ahuja et al., 1993a, Edmonds and Karp, 1972] on the flow network RFN,

$$\hat{\mathcal{T}} = \arg\min_{f(\cdot)} \sum_{e\in E} w(e) \cdot f(e), \tag{4.6}$$

where $f(e)$ is the amount of flow on every edge e.

Generally, solving a minimum-cost flow problem requires the pre-knowledge of the amount of flow sending from source to sink, denoted as k. Here k is set as a comparably large value (e.g. the total number of reads), the edge connecting *source* to *sink* will consume the extra amount of flow beyond the number of transcript copies which flow through the read vertices.

4.5 Acceleration with compressed flow network

The time complexity of the algorithms solving the minimum-cost flow problem is $O(|V|^3)$ [Ahuja et al., 1993b, Goldberg and Tarjan, 1989, Orlin, 1997], $|V|$ is the number of vertices. Given the size of reads, which could be in the order of millions, the problem can be intractable. Here we introduce a heuristic to compress the read flow network into a much smaller network with minimal loss of accuracy. The idea is to remove highly repetitive reads in high coverage region by clustering these reads into groups while still retaining the relationships among them. Given a compression parameter γ, the vertex set V of the flow network can be partitioned into a set of clusters $\Pi = \{\pi_1, \pi_2, \cdots, \pi_c\}$ of V, such that the reads within each cluster contain consistent splice junctions; have homogenous out-going edges and differ at most γ bases at both boundaries (Figure 4.3).

1. *Vertex homogeneity.* $\forall v_1, v_2 \in \pi_i$, $\pi_i \in \Pi$, v_1 and v_2 are either both the 5' end reads of one fragment or both the 3' end, and v_1 and v_2 either have the same set of splice junctions in their alignments or have no splice junctions;

2. *Edge homogeneity.* $\forall v_1, v_1' \in \pi_i$, $\forall v_2, v_2' \in \pi_j$, $\pi_i, \pi_j \in \Pi$, there exists no edge between v_1 and v_1' or between v_2 and v_2', and the edges between v_1 and v_2 represent the same set of paths in the splice graph as the edges between v_1' and v_2';

3. *Alignment adjacency.* $\forall v_1, v_2 \in \pi_i$, $\pi_i \in \Pi$, the 5'-most base of v_1 is at most γ bases away from that of v_2 on the genome, and the same for their 3'-most

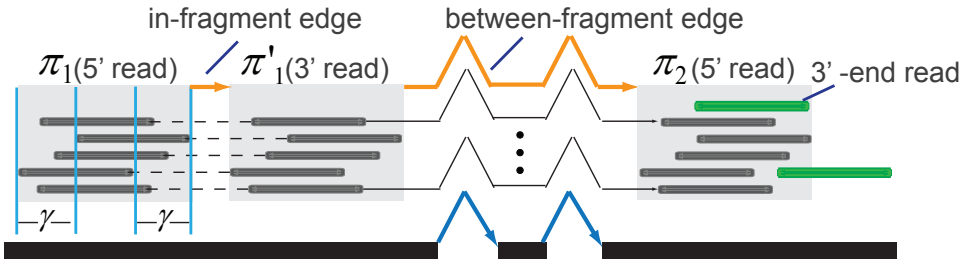


Figure 4.3: An example of the compressed flow network. Reads colored black are grouped into 3 clusters (light gray). Edges connecting the reads in the original RFN are collapsed into two edges (colored orange) in the compressed network. The two reads colored green cannot be clustered into π_2 because they violate the vertex homogeneity and alignment adjacency, respectively.

bases.

In this way, the vertex set V can be reduced to the set of vertex clusters Π, and the duplicated edges between vertices of two clusters can be removed if they represent the same path in the splice graph. The distances of duplicated edges may differ by at most 2γ bases, but the minimum weight of all the duplicated edges will be assigned to the only edge kept in the compressed flow network. The capacity of each vertex changes to the size of the cluster, and the capacity of an edge is the number of duplicated edges in the original flow network. Therefore, γ adjusts the degree of heterogeneity of reads in each clusters.When γ goes larger, generally more reads on the same exon can be grouped into one cluster and more reads containing the same splice junctions can also be clustered together. As a result, the compressed flow network has less vertices and edges and its size will become closer to that of the splice graph. In practice, γ is set to half the expected fragment length, $\gamma = len(r)/2$, which improves speed by significantly reducing the size of the flow graph while retaining high accuracy by allowing sufficient overlaps among reads in a cluster.

The calculation of partition can finish in $O(|V|)$ time. Our simulation studies have demonstrated that the compression may greatly reduce the time cost while maintaining a satisfactory accuracy of the assembled transcript copies.

4.6 Consolidating transcript reconstruction across alternative splicing events

The alternative splicing events (ASEs) happening between two cluster vertices will lead to more than one ways to connect them. In presence of multiple ASEs, it is important to avoid a simple enumeration of all possible isoforms from the combinations of variants in the ASEs. Therefore, we leverage the reads that span multiple ASEs to help evaluate the likelihood of existence of a possible isoform, using the MultiSplice features developed in our previous work [Huang et al., 2012]. Formally, a MultiSplice is a sequence of adjacent exons on a path of the splice graph, such that reads longer than a particular length may span all these exons. These features are calculated and incorporated here to reduce the possibility of linking the vertices into false transcripts.

Let e denote an edge in the compressed flow network. Let b denote the MultiSplice feature that consists of the same set of exons as the path indicated by e. Let $\psi(b)$ denote the size of the sampling window of b [Huang et al., 2012], which is the number of positions that a read could fall on in order to cover all exons of b (Figure 4.4). If no read is observed spanning b, the existence of edge e cannot be confirmed. In this case we assign a penalty to the weight of e by calculating the probability of observing

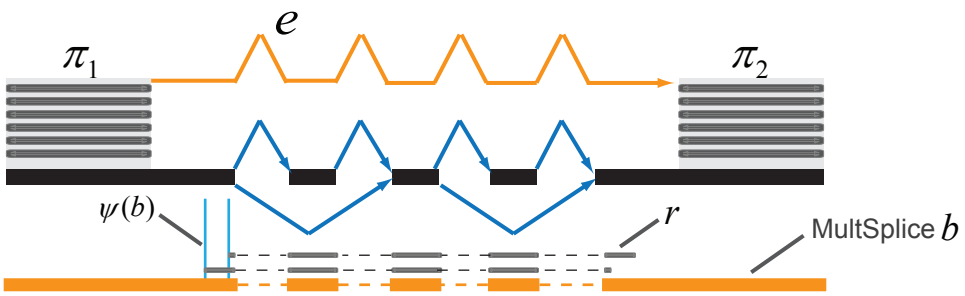


Figure 4.4: An example of a MultiSplice feature. Two ASEs (both are exon-skipping) reside between the clusters π_1 and π_2. The feature b consists of 5 exons on the path indicated by edge e. Two possible alignments of read r are shown in order for r to span b and confirm the existence of edge e. The possible positions of such alignments then give a sampling window of b (the window bounded by the two light blue lines).

no spanning read,

$$P(e \text{ not confirmed}) = \left(1 - \frac{\psi(b)}{len(e)}\right)^{|c_e|}, \tag{4.7}$$

where c_e is the capacity on e. If $\psi(b) = 0$, no read may span b at the given read length, $P(e \text{ not confirmed}) = 1$. Thereafter, we adjust the weight of e by adding $-\log P(e \text{ not confirmed})$ to $w(e)$.

4.7 Experimental results

We compared the performance of our method *Astroid* with four other state-of-the-art approaches for transcript reconstruction, including two "genome-guided" methods [Garber et al., 2011b] with different heuristics, Cufflinks [Trapnell et al., 2010b] and Scripture [Guttman et al., 2010] (Cufflinks 2.0.2 and Scripture beta version 2 were downloaded, Cufflinks was run in the mode that carries out both reconstruction and quantification and without -g/-G option), one representative method for Lasso-based "genome-guided" assembly IsoLasso [Li et al., 2011b] (IsoLasso version 2.6.0) and one "genome-independent" approach Trinity [Grabherr et al., 2011] (Trin-

ity version 2012-10-05). The assembled transcripts from Trinity were generated in fasta format and were mapped to the reference genome using BLAT [Kent, 2002] with default parameters. Only hits with complete match were considered in the comparisons. To understand the sensitivity and specificity of the transcript reconstruction as well as the accuracy of transcript quantification, we first did comparison on all five approaches using simulated datasets of varying sampling depths. We then compared the genome-guided assembly methods (Trinity excluded) on two real RNA-seq datasets, MAQC data[Shi et al., 2006] and Alexa-seq data[Griffith et al., 2010], where qPCR of a subset of transcripts are available to assess the accuracy of quantification using RNA-seq.

4.8 Simulation Studies

4.8.1 Data Simulation

We developed a simulator that mimics a real RNA-seq experiment and generates fragments from provided transcript copies. The simulation process consists of three steps: (1) Build a synthetic transcriptome by randomly assign copy numbers to all the genes and isoforms in the annotation database and set this as the true profiles. (2) Randomly cut the transcripts in the synthetic transcriptome into small fragments and dynamically check the lengths of the generated fragments. Fragments with lengths in a certain range (e.g. $[150bp, 350bp]$) are selected with probability to construct the sequencing library. This step stops when the number of fragments in the library exceeds the pre-specified sequencing depth. (3) $2 \times 75bp$ paired-end reads are sampled

from both ends of these selected fragments.

4.8.2 Matching Criteria

We evaluate the assembly results using similar criteria proposed in IsoLasso [Li et al., 2011b]. The assembled transcripts are compared with all the expressed transcripts in the profile (referred as "reference transcripts"). Two multi-exon transcripts are considered matched if they satisfy that (1) they contain the same set of exons; (2) all the exon boundary coordinates are identical except the start of the first exon and the end of the last exon. Also, two single-exon transcripts match if and only if at least 50% of the exons are overlapped. We adopted sensitivity and precision to measure the accuracy of the assembly results. Let M denote the number of reference transcripts. N out of M' assembled transcripts can be matched to the reference transcripts. Hence, $sensitivity = \frac{N}{M}$, and $precision = \frac{N}{M'}$.

4.8.3 Quantification Accuracy Criteria

Both Cufflinks and IsoLasso quantify transcript expression in the unit of FPKM. In Astroid, we use eTPM. However, these measurements cannot be directly compared. Therefore, we evaluate the quantification accuracy by the correlation between the transcript abundances estimated by each method and the true profiles. Pearson correlation [Bohnert and R., 2010] is adopted for this assessment. Let Y denote the true copy numbers of the transcripts and $\hat{Y}$ denote the estimated abundance. The correlation is calculated as $r(Y, \hat{Y}) = cov(Y, \hat{Y})/(\sigma_Y \cdot \sigma_{\hat{Y}})$, giving a value between -1 and $+1$. Higher correlation indicates more accurate estimation results.

4.8.4 Results

We conduct our first experiment to compare the performance of different methods on the transcriptome level. 30 million 2×75bp paired-end reads (insert size around 250bp) were simulated from the human transcriptome using RefSeq transcripts annotation. According to the profile, 18,374 transcripts from 13,030 genes were expressed. The reconstructed full-length transcripts of each method were matched against the ground truth, then the sensitivity and precision were assessed against different gene expression quantiles.

As shown in Figure 4.6 (a), Astroid consistently acquired highest sensitivity with increasing gene coverage. Even for the lowly expressed genes (bottom 10%), Astroid successfully recovered around 95% of these transcripts which is more than at least 20% of all the other methods. The precision of Astroid also outperformed the others on the bottom genes (shown in Figure 4.6 (b)). As gene expression climbs, the precisions became comparable between Astroid and Cufflinks, but were smaller than that of IsoLasso. This is probably related to the shrinkage strategy taken by IsoLasso which eliminates a large portion of transcripts through Lasso [Tibshirani, 1996].

Figure 4.6 (c) illustrates the quantification accuracy of each method. Astroid achieved highest correlation across different gene expression and demonstrated its ability of highly precise quantification through directly assembling transcript copies. However, both Cufflinks and IsoLasso showed very poor estimation. A further investigation on Cufflinks and IsoLasso abundance estimation results revealed that they both provided extremely high FPKM for short transcripts (less than 300bp) which

Table 4.2: Summary statistics of each method with various sampling depths. Correlation values in parentheses are calculated on only long transcripts (length > 300bp.

Methods	sensitivity		
	10M	20M	30M
Astroid(γ=0)	79.28%	91.71%	94.30%
Astroid(γ=30)	79.20%	91.76%	94.22%
Astroid(γ=50)	79.08%	91.64%	93.87%
Cufflinks	49.43%	74.48%	81.50%
IsoLasso	2.86%	23.97%	45.83%
Scripture	38.51%	62.13%	74.04%
Trinity	3.36%	13.01%	23.04%
Methods	precision		
	10M	20M	30M
Astroid(γ=0)	51.44%	80.23%	86.61%
Astroid(γ=30)	51.31%	80.01%	86.28%
Astroid(γ=50)	51.18%	79.47%	85.78%
Cufflinks	51.31%	75.75%	79.55%
IsoLasso	19.83%	75.26%	85.81%
Scripture	12.46%	26.34%	39.74%
Trinity	1.74%	6.13%	12.32%
Methods	correlation (long transcripts only)		
	10M	20M	30M
Astroid(γ=0)	.805 (.801)	.870(.868)	.922(.919)
Astroid(γ=30)	.808(.805)	.872(.869)	.918(.914)
Astroid(γ=50)	.808(.806)	.874(.870)	.912(.919)
Cufflinks	.106(.631)	-.033 (.773)	-.018(.808)
IsoLasso	-.027 (.356)	0.116 (.559)	.011 (.755)
Scripture	N/A	N/A	N/A
Trinity	N/A	N/A	N/A

is quite inconsistent with the profile. Similar observation was also reported by Li, *et al.* [Li and Dewey, 2011]. Excluding the abnormal results on these short transcripts, the correlation increases for both methods, but still falls behind Astroid. Astroid, however, was not heavily affected by the length of transcripts because of its capability to explicitly model the distance between reads and transcription start and termination sites.

We also look into the effect of the compression parameter on Astroid. According

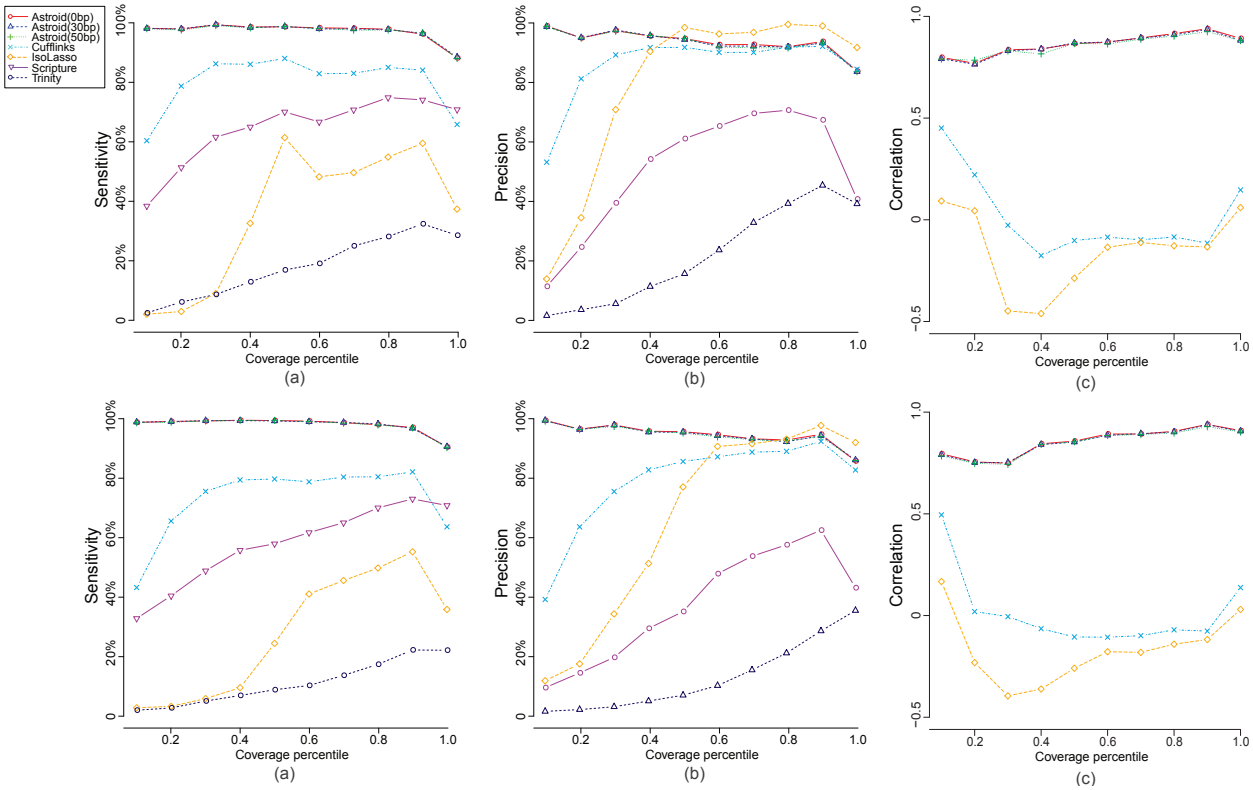


Figure 4.5: Performance comparison by Astroid with 3 different compression parameters (0bp, 30bp and 50bp), Cufflinks, IsoLasso, Scripture and Trinity on the synthetic human transcriptome dataset. (a), (b) and (c) are the sensitivity, precision and correlation (excluding Scripture and Trinity) against increasing gene coverage when the mean sequenced fragment length is 350bp. (d), (e) and (f) are the sensitivity, precision and correlation (excluding Scripture and Trinity) against increasing gene coverage when the mean sequenced fragment length is 450bp (The legends of these three subfigures are the same as (a), (b) and (c), respectively).

Table 4.3: Computational performance on there 30M 2×75bp paired-end datasets with different mean fragment lengths. All programs were run on an Intel Xeon E5-2450 32-core 2.10 GHz Linux server with 98GB of RAM.

Methods	250bp	350bp	450bp
Astroid($\gamma = 0$)	$24h$	$10h$	$4h$
Astroid($\gamma = 30$)	$7h$	$6h$	$1h$
Astroid($\gamma = 50$)	$1h$	$1h$	$40min$
Cufflinks	$40min$	$38min$	$30min$
IsoLasso	$10min$	$5min$	$4min$
Scripture	$15min$	$16min$	$18min$
Trinity	$8h$	$7h$	$6h$

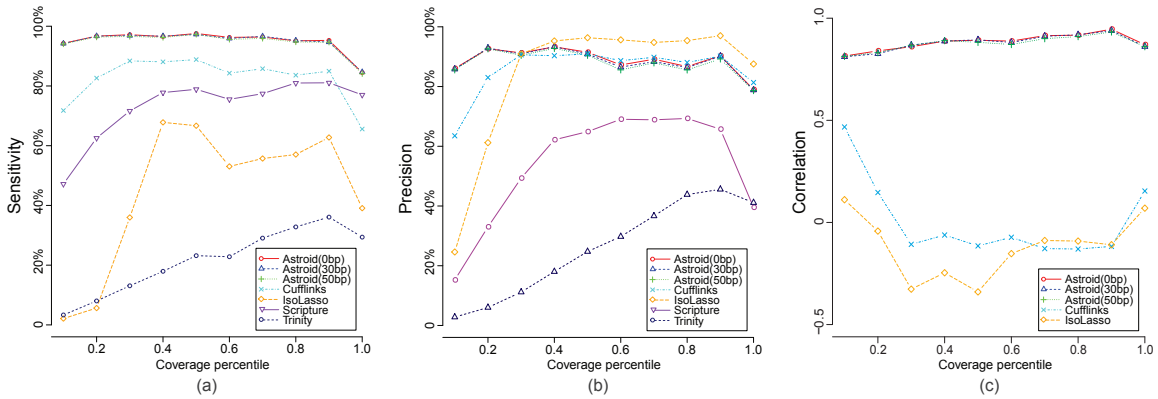


Figure 4.6: Performance comparison of Astroid with 3 different compression parameters (0bp, 30bp and 50bp), Cufflinks, IsoLasso, Scripture and Trinity on 30M 2×75bp (insert size around 250bp) paired-end dataset. Evaluation measurements were plotted against different gene expression quantile (in 10% increments). (a) Each point in the plot represents the sensitivity of one method which is the ratio between the number of matched transcripts and the reference transcripts within one quantile. (b) Each point represents the precision of one method which is the number of matched transcripts and the total assembled transcripts within one quantile. (c) The correlation of transcript expression is computed on the set of matched transcripts for each method.

to the results shown in Figure 4.6, we do observe that Astroid baseline ($\gamma = 0bp$) performs better than the other two with positive γ, but the difference is not that significant. Meanwhile, as shown in Figure 4.3, the time cost improves from 1 day to 1 hour as γ increases from 0bp to 50bp. This suggests that significant improved in efficiency can be achieved without much degradation of its performance. Therefore, in real practice, we may set the compression parameters at a comparably larger value ($\gamma = len(r)/2$). We use this setting in real data experiments.

We next evaluate how the sampling depth may affect the performance of each method. To do this, we first sample 10M and 20M 2×75bp paired-end reads by random selection out of the 30M dataset. Table 4.2 shows the overall sensitivity, precision and correlation on these three datasets. From the statistics, we see that both

the sensitivity and precision improve for all methods as more reads are sequenced. Apparently, higher sampling depth is more conducive for inferring transcript structures. Similar with previous observation, Astroid showed best performance against various sampling depths, which indicates that eTPM computed from the effective transcripts, is a robust measure for estimating the relative transcript abundance.

Table 4.4: Summary statistics on two 30M 2×75bp paired-end datasets, with mean insert size of 350bp and 450bp respectively.

Methods	350bp		
	sensitivity	precision	correlation (long)
Astroid($\gamma = 0$)	95.23%	88.85%	.923(.919)
Astroid($\gamma = 30$)	95.18%	88.60%	.919(.915)
Astroid($\gamma = 50$)	95.05%	88.52%	.914(.910)
Cufflinks	78.53%	81.97%	0.089(.682)
IsoLasso	37.47%	81.04%	.011(.730)
Scripture	64.97%	34.19%	N/A
Trinity	19.61%	9.60%	N/A
Methods	450bp		
	sensitivity	precision	correlation (long)
Astroid($\gamma = 0$)	95.48%	89.97%	.929(.925)
Astroid($\gamma = 30$)	95.47%	89.83%	.927(.923)
Astroid($\gamma = 50$)	95.35%	89.69%	.923(.919)
Cufflinks	71.59%	75.57%	.086(.601)
IsoLasso	28.95%	68.75%	−0.022(.618)
Scripture	58.14%	23.76%	N/A
Trinity	11.85%	5.25%	N/A

Besides sampling depth, we also investigate how varying fragment lengths would affect the performance of each method. Table 4.4 show the summary statistics on two 30M paired-end datasets with mean insert size of 350bp and 450bp, respectively. Combining with the results shown above on the 30M dataset with mean insert size of 250bp, we see that the performance of Astroid, especially the precision, improves

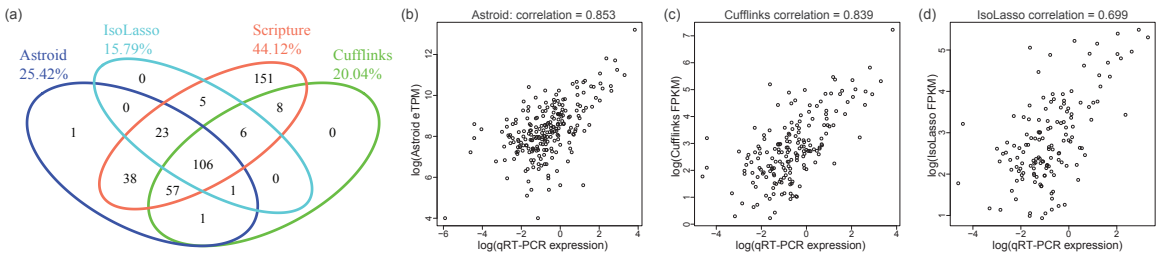


Figure 4.7: (a). Venn Diagram of qRT-PCR validated transcripts reconstructed by Astroid, Cufflinks, Scripture and IsoLasso. (b)-(d) Scatter plots (on log_e scale) of transcript abundance estimated by Astroid, Cufflinks and IsoLasso, respectively, against qRT-PCR expression on the set of qRT-PCR validated transcripts that are reconstructed in full length by each method.

as the the fragment length increases. This demonstrates Astroid's capability of taking advantage of longer fragments to infer the true set of transcripts. Moreover, the correlation computed between the true profiles and the estimated abundance of the matched transcripts of Astroid is much higher than those of Cufflinks and IsoLasso, suggesting that the eTPM estimated by Astroid provides a more accurate quantification of transcript quantities. Figure 4.5 also demonstrated Astroid's superior performance over the other existing methods on varying fragment lengths datasets.

4.9 Experiments with real RNA-seq datasets

4.9.1 MAQC data study

For evaluation on real RNA-seq experiments, we first compared the four genome-guided transcript reconstruction approaches Astroid, Cufflinks, Scripture and IsoLasso using the RNA-seq dataset from Microarray Quality Control (MAQC) project Human Brain Reference (HBR) sample [Shi et al., 2006] (NCBI Short Read Archive accession number SRA012427). This dataset contained 23 million 2x50bp paired-end

reads generated from three lanes. Besides RNA-seq data, 907 transcripts were analyzed with TaqMan qRT-PCR for their expression, including 893 that could be matched to RefSeq transcript annotation [Pruitt et al., 2008] (accession number GSE5350). We focused our analysis on this subset of validated transcripts.

Among the 893 qPCR-validated transcripts, Astroid correctly reconstructed 227, with a sensitivity of 25.42% (227 out of 893). This sensitivity is higher than those of Cufflinks (20.04% or 179 of 893) and IsoLasso (15.79% or 141 of 893). This demonstrates Astroid's good ability to reconstruct full-length transcripts. The venn diagram shown in Figure 4.7 (a) illustrated a good consistency of their assembly results. We notice that Scripture reconstructed the most number of validated transcripts (44.12% or 394 of 893). This is due to the strategy of Scripture which tries to enumerate all possible transcripts given the exons and junctions observed from RNA-seq data. This strategy may highly increase the sensitivity but it also introduces large amount of false positives, especially on the genes with high coverage. In fact, the total number of assembled transcripts is 92,977 for Scripture which corresponds to a precision of 0.42% (394 of 92,977) and it is only half of Astroid (0.87% or 227 of 26,119). However, it is surprisingly that both Cufflinks and IsoLasso showed lower precision than Scripture on the identification of validated transcripts: 0.26% (179 of 69,011) and 0.10% (141 of 135,085). A close examination revealed that majority of their reconstructed transcripts are very short single-exon transcripts with low coverage, which are probably just background noises due to sequencing or mapping biases.

Next, we examined the transcript expression measured by qRT-PCR experiments and the expression estimated by each method (excluding Scripture) on the set of

transcripts that were validated and correctly reconstructed: scatter plots are shown in Figure 4.7 (b)-(d). Transcript abundance inferred by Astroid reached a Pearson correlation as high as 0.853 on all the transcripts it correctly assembled, slightly higher than Cufflinks (0.839) and much higher than IsoLasso (0.699).

This result demonstrated that Astroid is competitive for transcript quantification. We further ran Cufflinks in its quantification-only mode by providing the RefSeq transcript annotation. The estimated transcript abundance by Cufflinks on all 893 validated transcripts had a Pearson correlation of 0.866, consistent with its previous reports on MAQC dataset for transcript quantification [Roberts et al., 2011]. The difference between Cufflinks without transcript annotation and with annotation suggests that downstream analysis such as transcript quantification can be significantly altered by transcript reconstruction results. On the other hand, Astroid shows the prominent ability of discovering the underlying transcripts and providing reliable expression estimates simultaneously.

4.9.2 Alexa-seq data study

We further applied Astroid and other methods to a real RNA-seq dataset used by Alexa-seq [Griffith et al., 2010], an alternative expression/transcription analysis method. Total 262 million Illumina paired-end RNA-seq reads (36bp or 42bp) were generated from two cell lines: fluorouracil (5-FU)-resistant and -nonresistant human colorectal cancer cell lines, MIP101 and MIP/5-FU. The raw RNA-seq reads were downloaded from Alexa-seq website (`http://www.alexaplatform.org/alexa_seq/`). 167 million paired-end reads were generated from MIP101 sample and 89.82%

Table 4.5: Summary statistics on the validated set of exons.

Methods	# exons reconstructed		correlation with qRT-PCR expression	
	MIP101	MIP/5-FU	MIP101	MIP/5-FU
Astroid	137	114	0.99	0.81
Cufflinks	124	66	−0.02	−0.03
IsoLasso	131	76	0.99	0.87
Scripture	105	60	N/A	N/A

of them were successfully mapped by MapSplice using human hg18 reference genome. The rest 95 million reads came from sample MIP/5-FU, among which 90.26% were mapped by MapSplice. Alexa-seq also provided qRT-PCR validation on 192 alternatively expressed exons. We focus the comparison of all the methods on identification of all the validated exons. One exon is considered reconstructed by one method if: (1) at least one assembled transcript contains this exon; (2) both boundaries of the identified exon have to match the hg18 annotation unless this exon is transcription start/end; (3) if the exon is transcription start/end, only downstream/upstream boundary of this exon is required to match the annotation, respectively. The estimated abundance on this exon is collected as the cumulative estimated abundance on the exon of all the transcripts assembled covering it.

Table 4.5 shows the number of validated exons successfully reconstructed and the correlation between the estimation and the qRT-PCR expression by each method. From the results, we observe that Astroid reconstructed the highest number of exons in both samples among all the assembly tools. This suggests that Astroid successfully reconstructed the transcripts containing these target exons. Meanwhile, correlation between estimated abundance and qRT-PCR expression was computed on the set

of reconstructed exons by each method. Astroid and IsoLasso acquired the highest correlation (0.99) on sample MIP101, much higher than Cufflinks (-0.02). The correlation by Astroid dropped on MIP/5FU sample, but was still comparable to IsoLasso, which also outperformed Cufflinks.

Although Astroid consistently performs better than the other methods on the two real RNA-seq datasets, it is noticed that its improvement is not as significant as that in simulation experiments. After further investigation, we found that: (1) for real datasets, we only have access to a very small set of validated transcripts or exons supported by abundant read alignments. But for simulation, we sampled reads from the whole transcriptome containing genes with a large dynamic range in their expression. The splice junctions with relatively low read support tend to be filtered out by methods like Cufflinks and IsoLasso, which lead to their failure in reconstructing the correct set of full-length transcripts; (2) for MAQC dataset, the transcripts with PCR validation are mainly from single-isoform genes. As we know, it is easier to reconstruct and quantify single-isoform genes than multi-isoform genes. As a result, the differences among these methods are minimal.

4.10 Discussion

In this chapter, we have presented a novel method Astroid for simultaneous transcript reconstruction and quantification. Compared with existing methods which typically reconstruct isoforms in a splice graph, our approach provides a unique solution by piecing individual reads into a set of effective transcript copies. A novel measure for transcript abundance eTPM has also been defined based on the assembled effective

copies, rather than indirect estimators that fully depend on the read count. The problem of the reconstruction of effective transcript copy has been modeled as a minimum-cost flow problem, which allows the solution of a maximum-likelihood set of copies.

We evaluated Astroid as well as four existing methods using both simulated data and real data. In general, the eTPM measure generated by Astroid has a better overall correlation with the ground truth or qRT-PCR measurement than FPKM output from Cufflinks and Isolasso. However, further validations using real datasets are still necessary in checking out the relationships among eTPM, TPM and FPKM in terms of their accuracy in inferring the abundance of alternative transcripts in multi-isoform genes as well as reconstructing isoforms of genes with relatively low expression. We are also interested in validating whether eTPM or TPM would be able to effectively normalize transcript abundance by the size of transcript library that is sample-specific, alleviating the risk of comparing transcriptoms with drastically different transcript composition.

Our approach is built on the assumption that short read sequencing may only capture a fraction of each mRNA molecule. Hence, the sampling "gaps" on transcripts that we have modeled has the potential to handle uneven read distribution due to various biases, such as positional bias and sequence bias. Ill-formed copies which contain only a proportion of the expected transcript may indicate an aberrant distribution of the observed reads and suggest possible biases. For example, if 3' end positional bias is observed, we may compensate the less sequenced 5' end by allowing a larger gap between 5' end and one fragment. We are currently working on potential

methods to correct these biases within the existing framework.

Chapter 5 Transcriptome analysis on large-scale RNA-seq datasets

5.1 Introduction

The RNA-seq technologies which sequence the mRNA transcriptome at an unprecedented level have enabled a more comprehensive analysis of the presence and quantity of the mRNA transcripts in the transcriptome. However, only examining one individual or limited number of samples may not reveal the overall picture of cell functioning, development and differentiation. Take cancer study for example, every tumor is different. The sample size is the key to understanding of the genetic mechanisms behind various diseases, and therefore crucial for precise disease diagnosis, prognosis and treatment.

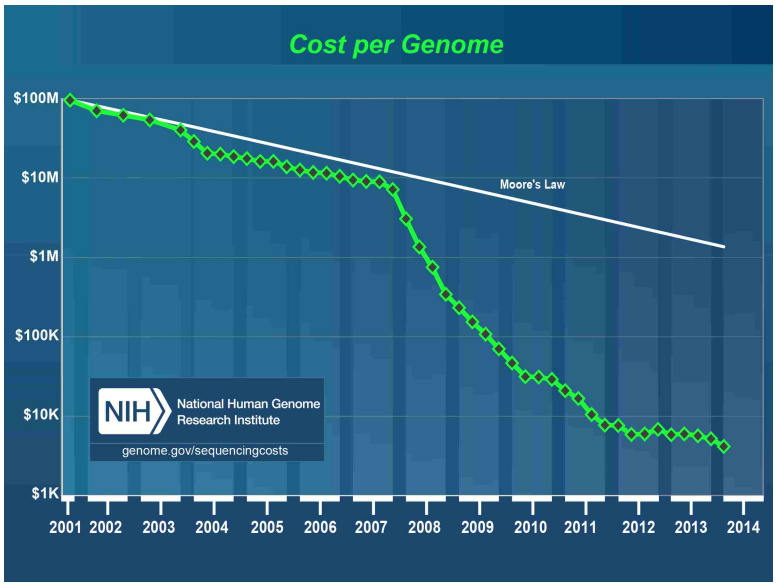


Figure 5.1: Typical cost of sequencing a human-sized genome, on a logarithmic scale. The curve decreases much faster than Moore's law [Moore, 1965]. Figure accommodated from Wikepedia(`www.wikepedia.com`)

Nowadays, with the rapid development of sequencing technology, we are able to sequence a sample or tissue with much cheaper cost, which directly leads to the capability of sequencing much more samples with same budget. Figure 5.1 illustrates that now the cost of sequencing a human-sized genome has dropped to $\$100,000$ by 2009 and $\$1,000$ by 2014. Empowered by this advancement, several large-scale genome sequencing projects have launched to accelerate the understanding of the molecular basis of human diseases, especially cancer, such as ICGC (International Cancer Genome Consortium), TCGA (The Cancer Genome Atlas) and CCLE (The Cancer Cell Line Encyclopedia). For example, the RNA-seq data of TCGA consists of approximately 5000 samples, more than 1 trillion reads and approximately 50 TeraBytes of binary files. The *big data* have brought great challenges of insight discovery and process optimization beyond computing and storage resources. In this chapter we will elaborate the computational difficulties and introduce the standard Cufflinks+Cuffmerge pipeline for processing massive RNA-seq data along with its problems. Lastly, we will present an alternative workflow which aims at resolving the issues with previous approaches. It performs comparative transcriptome analysis across hundreds of RNA-seq samples and efficiently discovers the biological signatures.

5.1.1 Computational challenges in studying cancer transcriptomes

Heterogeneity between tumors and even within tumor has been hindering therapy [Hiley et al., 2014, Fisher et al., 2013, Junttila and de Sauvage, 2013]. To understand the mechanism of cancer in a molecular level, cancer consortiums such as TCGA [TCG] and ICGC [ICG] have generated a comprehensive set of next generation sequenc-

ing (NGS) data on thousands of patient samples. With whole genome sequencing data, much progress has been made on uncovering genomic mutations that drive tumor evolution [Network, 2012, 2013b, Network et al., 2013]. On the other hand, the RNA-Seq data captures the snapshots of the entire transcriptomes with hundreds of millions of short reads per sample in an unprecedented depth and resolution. The alignments of these reads has made possible the discovery of novel splice variants, gene fusions and expression signatures, among others. To date, the analyses on these transcriptomes have largely been focused on gene expression patterns associated with cancer subtypes [Network et al., 2013, Hoadley et al., 2014, Network, 2013b,a]. Another complexity yet remaining barely explored has been the regulation of alternative splicings that differentiate the transcriptomes, which directly determines the variety and quantity of transcript isoforms and ultimately the proteins. Disentangling this transcription-level heterogeneity could further highlight misregulated or aberrant isoforms, especially those implicated as determinants of cancer subtype by exhibiting subtype-specific presence or usage.

However, to compare hundreds of transcriptomes is not a trivial extension from existing differential expression/transcription analysis approaches. One specific challenge in characterizing cancer transcriptome is the ability to discover and catalogue novel cancer-specific events, many of which may be rare and may not be curated into existing annotations. A common practice in expression study, such as the pipeline that processed most TCGA samples, feeds the reads into a reference transcriptome-based program like RSEM [Li and Dewey, 2011] to quantify known isoforms (Figure 5.2b). The estimated isoform abundance is then carried over to downstream analyses such

as subtype classification and survival prediction. Unfortunately, this approach is only confined to the isoforms within the reference transcriptome and simply ignores aberrant transcription events.

To avoid the limitation of incomplete annotation, one often opts to use an ab initio isoform reconstruction method to first infer isoforms from raw read alignments. Tools such as Cufflinks and Scripture can be leveraged with partial or no help of a-priori annotation. While this is promising for small data sets, a caveat could be the scalability. Isoform reconstruction algorithms heavily depend on read alignments, gapped alignments that reveal splice junctions in particular, which are not noise free due to the length of the short reads coupled with the complexity of human genome. Although current aligners such as MapSplice and TopHat are capable of controlling the false junction discoveries, the set of total putative junctions can easily become intractable when considering all samples in the data set. In TCGA BRCA data, for example, on average the number of junctions in each sample is XXX with XX.

Another question remaining unsolved is how to use the power allowed by the sample size. The aforementioned approaches process each sample independently, from read alignment to isoform quantification. Given the wealth of the information provided in the large datasets with hundreds of samples, this approach may not be ideal. For example, splice junctions that are highly supported by read alignments in some sample may be poorly supported in others due to random sampling. Processing each sample independently may falsely eliminate any transcript missing a single splice junction, resulting incomplete transcriptome, complicating comparison as well.

Cufflinks is widely used tool for transcriptome analysis. It has a comprehensive

pipeline for transcriptome reconstruction, q In this chapter, we systematically investigate the effectiveness of Cufflinks pipeline for the joint analysis of massive RNA-seq data.

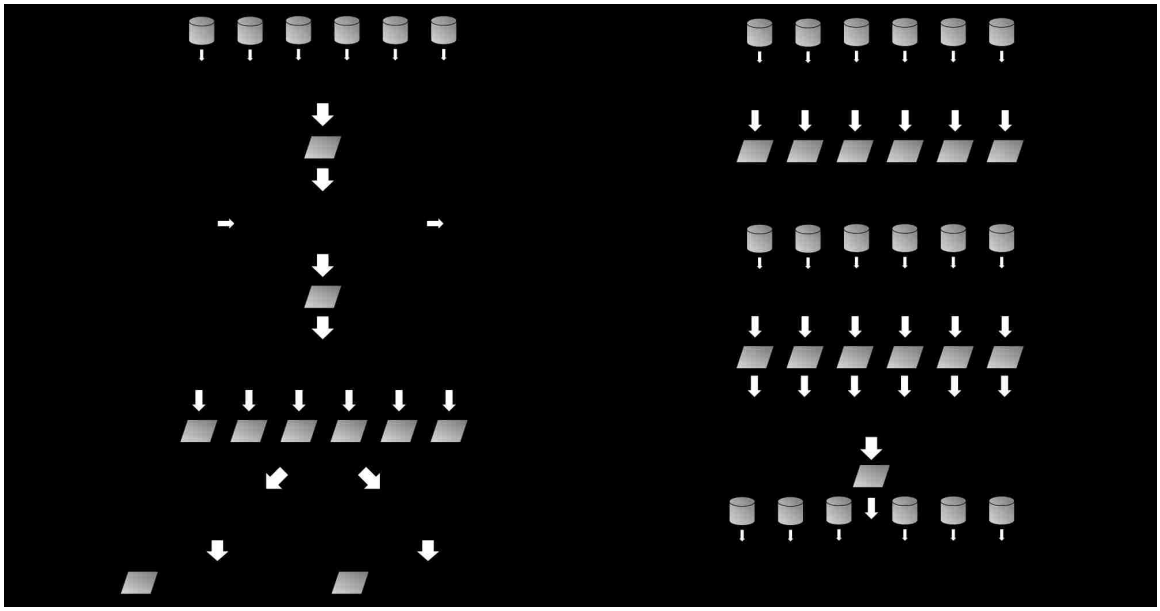


Figure 5.2: Overview of different workflows for differential analysis on large-scale datasets. a) Pipeline of proposed method. b) Typical TCGA pipeline guided by transcrptome annotation using RSEM. c) Standard Cufflinks+Cuffmerge pipeline optionally assisted by transcriptome annotation.

5.1.2 The TCGA breast cancer RNA-seq datasets

The Cancer Genome Atlas (TCGA) project was supported by National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) and its fundamental goal is to improve our ability to diagnose, treat and prevent cancer. Starting from 2006, it first aims at comprehensively mapping and characterizing the genomic changes in brain and ovarian cancers, as well as proving that a national network of researchers could effectively collaborate to generate large-scale genomic data

and make discoveries. The achievements further led to more resources devoted for characterization of more than 30 additional tumor types, including nine rare cancers.

The TCGA breast cancer (BRCA) consists of 819 RNA-seq samples. 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). All following analysis is performed on the TCGA BRCA RNA-seq dataset.

5.2 Existing pipelines on joint analysis of hundreds of cancer transcriptomes

Consider the complexity and scale of the TCGA BRCA dataset, current analysis usually rely on the prior knowledge of the annotated transcripts. There are mainly two trends of pipelines within this category and they differ in the read mapping stage. One first map the reads to the reference genome and then feed the read alignments to quantification models, such as Cufflinks with annotation mode to estimate the abundance of the reference genes/isoforms. The other one directly feeds the raw reads into reference transcriptome-based program like RSEM which uses built-in Bowtie to map the reads to the reference transcripts and then performs quantification. The abundance of the known isoforms is measured in RPKM/FPKM [Trapnell et al., 2010b] or TPM [Li et al., 2010a] and compared across samples. Both workflows are employed by many research groups and institutions, such as UNC Chapel Hill

and Broad institute. An obvious problem with this kind of approach is that by restricting the analysis on the reference transcriptome, aberrant genetic events will be overlooked, such as novel transcription and gene fusion.

To circumvent this problem, the other strategy doesn't depend on known transcript database, and uses Cufflinks+Cuffmerge pipeline for reconstructing a unified set of transcripts first (Figure 5.2c). Ideally, the "assembly-merge" model would work. But in real application, the ambiguity is greatly enlarged for merging transcripts from large number of samples and therefore largely complicates downstream analysis. In this section, we will systematically study this pipeline and show its limitations on large-scale data analysis.

Cufflinks is a comprehensive tool developed for mRNA transcriptome analysis using RNA-seq data. Provided with RNA-seq read alignments, it has several components: reconstruct isoform transcripts, quantify their abundances, and test for differential expression and regulation in RNA-Seq samples. Figure 5.3 illustrates a typical workflow of Cufflinks [Trapnell et al., 2010b]. Generally, when applied on multiple RNA-seq samples, Cufflinks first treats each sample independently and performs assembly and quantification analysis, Cuffmerge then collectively combines all individual results for further study.

5.2.1 Cufflinks running modes

Cufflinks is a powerful tool which provides multiple options allowing for user customized running modes. Table 5.1 summarizes two different assembly modes of Cufflinks: with and without guidance of annotated transcripts. The *RABT assembly*

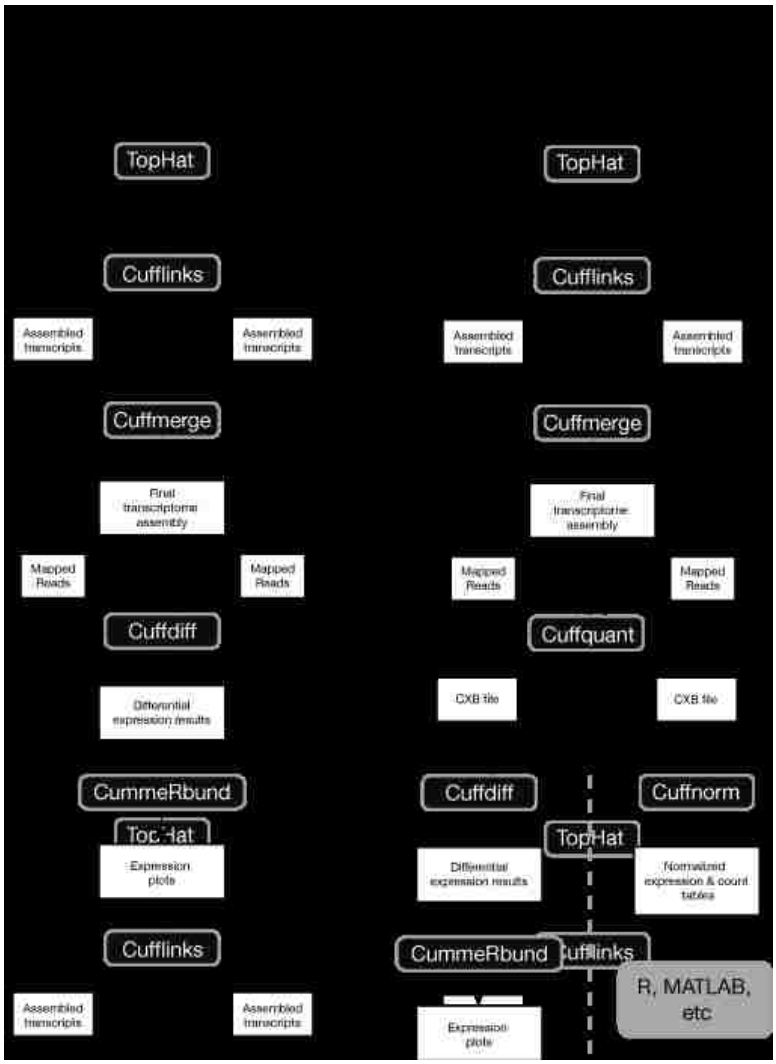


Figure 5.3: Overview of the Cufflinks workflow. The raw RNA-seq reads are first mapped to the reference genome using alignment tools such as Tophat [Trapnell et al., 2009c] or MapSplice [Wang et al., 2010a]. Cufflinks assembles the RNA-seq read alignments into a parsimonious set of transcripts, and then estimates the relative abundances according to a maximum likelihood model that assigns probability for each fragment to one transcript. In need of comparing multiple transcriptomes of different conditions, Cuffmerge is utilized for first merging the transcript assembly result of each individual sample. Following this, Cuffdiff takes input of the merged set of transcripts and the read alignments from all groups of samples, and tests for significant changes in gene/isoform expression, splicing and promoter use.

which refers to the annotation will include reference transcripts with no read coverage for completeness. Please note that "-G /ref.(gtf/gff)" is different from RABT assembly. It is quantification-only mode of Cufflinks which is used for estimate the

abundance of known gene isoforms. Table 5.2 summarizes four modes of Cuffmerge. Cuffmerge is designed to compare multiple assemblies, remove duplicates and automatically filter the transcripts that are probably artfifacts. Reference transcripts or/and reference DNA sequences may be fed to Cuffmerge to maximize overall assembly quality.

Table 5.1: Two modes of Cufflinks assembly. RABT assembly is guided by gene/isoform annotation. The reference transcripts are tiled with faux reads which are also combined with sequencing reads for transcript reconstruction. The assembled transcripts are further compared with the reference to determine whether they are sufficiently novel.

Mode Type	Option	Mode description
Cufflinks w/o annotation	N/A	Assemble transcripts solely from read alignments
Cufflinks assisted by annotation *RABT assembly*	$-g/ref.(gtf/gff)$	Reference transcripts are required to provide additional information

Table 5.2: Four modes of Cuffmerge. Reference transcripts or reference genome can be provided to Cuffmerge for guidance.

Mode Type	Option	Mode description
Cuffmerge	N/A	Merge transcripts solely from the input assemblies
Cuffmerge assisted by reference transcripts	$-g/ref.gtf$	The input assemblies are merged together with the reference GTF
Cufflinks assisted by reference sequence	$-s/ref.seq$	The input assemblies are merged guided by the reference genome
Cufflinks assisted by reference transcripts&sequence	$-g/ref.gtf\& -s/ref.seq$	Combine "-g/ref.gtf" and "-s/ref.seq"

5.2.2 Cufflinks investigation experiments

Using 819 RNA-seq datasets from TCGA BRCA project, we conducted analysis on Cufflinks pipeline. Following the workflow shown in Figure 5.3, we first ran Cufflinks assembly in both modes on each individual sample. Next, Cuffmerge in four modes

were applied on the two group of assemblies to merge the reconstruction results into a single gtf. Our first experiment is to compare two assembly modes of Cufflinks.

Comparison of Cufflinks without annotation and RABT assembly. Both Cufflinks without annotation and RABT assembly were ran on all 819 RNA-seq samples of TCGA BRCA project. We would like to compare the similarity of two modes on each individual sample. Cuffcompare was utilized for this task. As stated in Cufflinks manual, Cuffcompare "examines the structure of each the transcripts, matching transcripts that agree on the coordinates and order of all of their introns, as well as strand. Matching transcripts are allowed to differ on the length of the first and last exons, since these lengths will naturally vary from sample to sample due to the random nature of sequencing." It will generate a union of all transcripts from both inputs and report whether or not each transcript is present in the input assemblies.

Figure 5.4(a) summaries the count of isoforms from different origin in the Cuffcompare union set. We observe that the shared portion is much less than the unique ones for most samples and the RABT assembly mode generates much larger set of isoforms than the without annotation assembly. This result indicates that there exists large discrepancy between these two assembly modes. Figure 5.4(b) also supports this discovery. The percentage of shared isoforms is computed as: $\frac{\#\text{shared isoforms}}{\#\text{isoforms in Cufflinks w/o annotation}} \times 100\%$ and $\frac{\#\text{shared isoforms}}{\#\text{isoforms in RABT assembly}} \times 100\%$ for two modes, respectively. Clearly, RABT assembly produces significantly more isoforms than without annotation mode in most samples.

From these results, we found out that different modes of Cufflinks gave very

different results. Therefore, transcript level reconstruction remains very challenging and difficult.

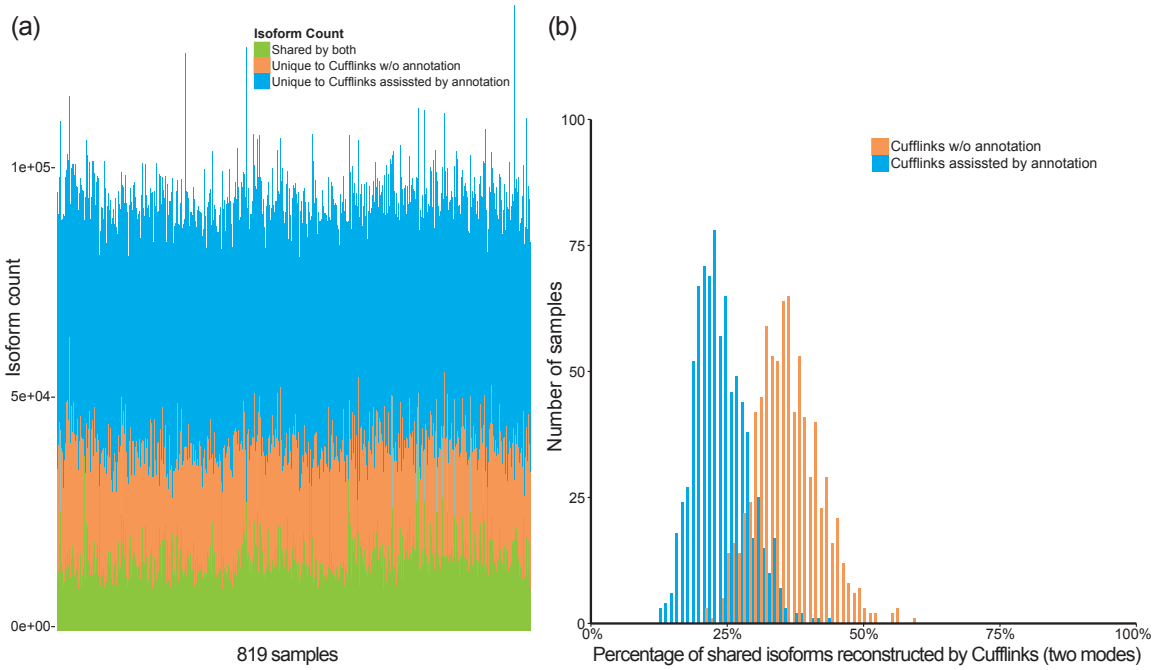


Figure 5.4: (a) For each sample, the counts of isoforms shared by both Cufflinks assembly modes and unique to only one mode are plotted. Each bar represents one sample and each color represents one origin of the isoforms. (b)Histogram of percentage of shared isoforms reconstructed by two Cufflinks assembly modes respectively.

Comparison between individual assembly and Cuffmerge result. Following Cufflinks assembly, Cuffmerge is applied to merge all assemblies into a single transcript set for further analysis. In this section, we would like to study the effectiveness of Cuffmerge. To achieve this goal, we combine the 819 RABT assembly results using Cuffmerge with "-gs" option and compare the merged result with each individual assembly. Because both RABT assembly mode and Cuffmerge with "-gs" option are assisted by references which leads to more power of filtering artifacts.

Figure 5.5 shows a histogram of percentage of shared isoforms in Cuffmerge result. We see the percentages are fallen in a very narrow range, approximately between

15.5% and 17.0%. This plot illustrates (1) very small portion of isoforms in the merged set matches to each individual sample, and (2) the percentages of shared isoforms are quite similar across the samples. Figure 5.8 also supports this finding which shows that there is a large number of isoforms in the merged set that aren't originated from one specific sample and Cuffmerge also throws away some isoforms it regards as "artifact" for every sample.

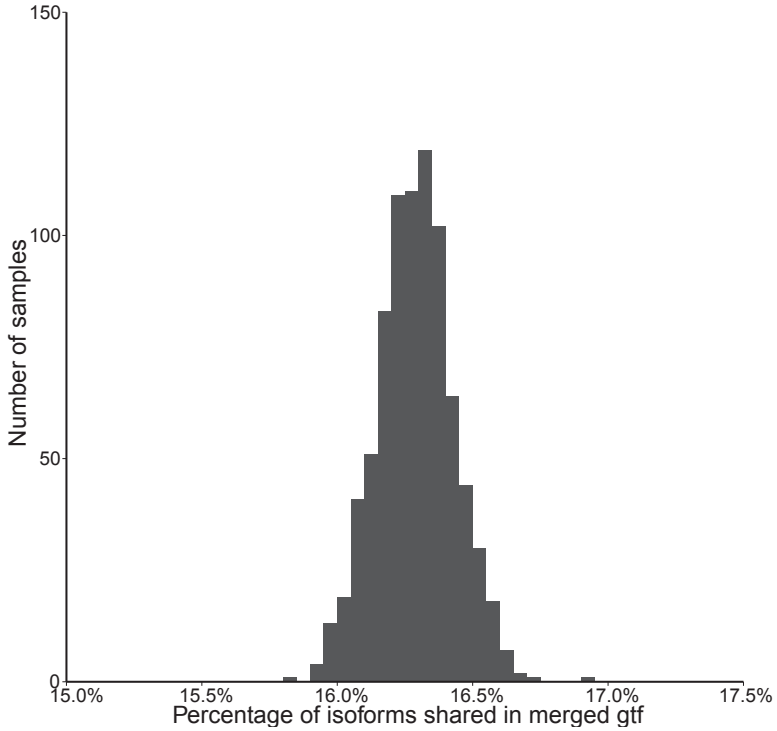


Figure 5.5: Histogram of percentage of shared isoforms in Cuffmerge result.

We further look into the isoform count composition within each sample (Figure 5.6a). Similar with previous observation, the numbers of shared isoforms with merget set are highly consistent across samples. Figure 5.6b shows a histogram of the percentage of isoforms discarded by Cuffmerge. For most samples, less than 2/3 of the isoforms are absent from the final merge set, probably due to some filtering based on the reference transcripts and reference genome.

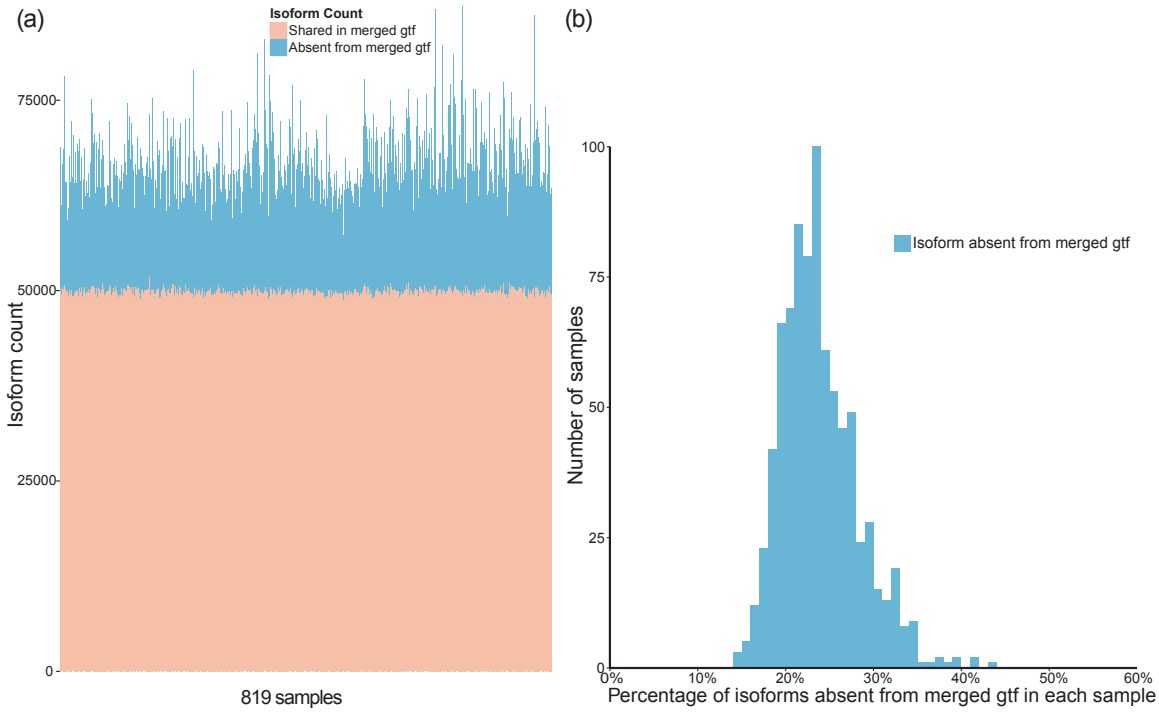


Figure 5.6: (a) For each sample, the counts of isoforms shared by by both individual sample and the merged set, and unique to only the sample (absent from the final merged result) are plotted. Each bar represents one sample and each color represents one origin of the isoforms. (b) Histogram of percentage of isoforms absent from Cuffmerge result in each sample.

Since the shared portions of isoforms are highly consistent across the samples, pairwise comparison is conducted on the shared transcripts between any two samples. The similarity distance between assemblies i and j is calculated as $\frac{\#\text{shared isoforms}}{\#\text{union isoforms of i and j}}$. Here we only consider the isoforms present in the merged set. The heatmap shown in figure 5.7 illustrates the average similarity between any two samples is around 0.9.

In conclusion, although individual assemblies may be quite similar between each other, uncertainty of transcript reconstruction in each sample can be largely amplified and propagated when looking at the entire data set: the merged set still has a lot of transcripts than one can not find within one individual sample.

Comparison of Cuffmerge results. Besides the study on each individual sam-

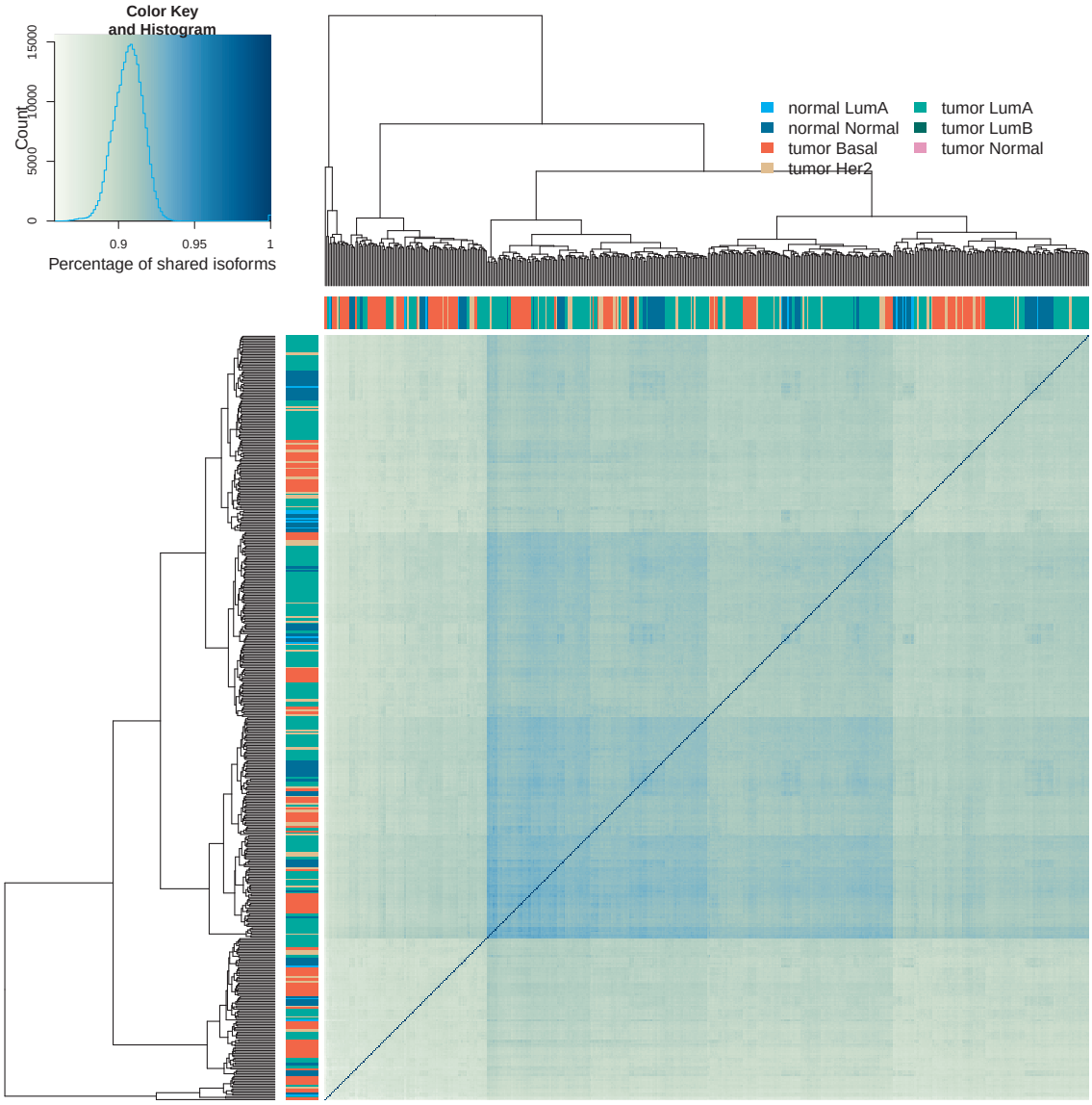


Figure 5.7: Heatmap comparing pairwise similarity among RABT assemblies. Only the shared isoforms with Cuffmerge result are considered. Darker color means higher similarity. Please note only 499 samples are included in the plot, since Cuffcompare is limited to 500 inputs and one is reserved for merged set.

ple, we further look into the consistency among the merged assemblies. Totally 8 different assemblies can be obtained from merging two groups of Cufflinks assembly. We denote them as: *merged*, *merged_opt_g*, *merged_opt_s* and *merged_opt_gs* for the merged results from Cufflinks assembly without annotation, and *rabt.merged* and other three with prefix "rabt." for Cufflinks RABT assembly. Here, "_opt" is adopted

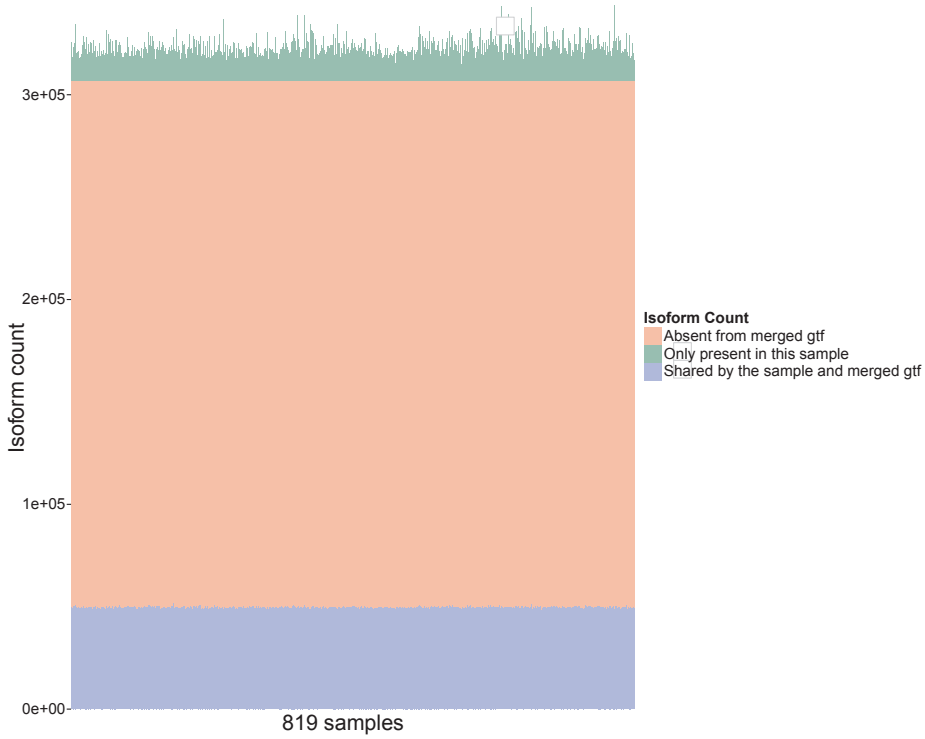


Figure 5.8: For each sample, the counts of isoforms shared by both individual sample and the merged set, and unique to either one are plotted. Each bar represents one sample and each color represents one origin of the isoforms.

to distinguish various Cuffmerge options. For example, "_opt_gs" indicates both reference transcripts and reference genome are provided.

Table 5.3 shows some summary statistics we collected from the merged assemblies and the UCSC hg19 reference transcripts set. From this table we observe that the number of genes and isoforms produced varies quite much among all 8 modes and they both differ significantly from the reference. Moreover, although it is mentioned in Cufflinks manual that the reference transcripts or reference genome can be introduced in Cuffmerge for filtering artifacts, the total number of isoforms generated are much higher when reference transcripts are fed. Besides this, the statistics between Cufflinks without annotation and RABT assembly are quite similar if of same

Cuffmerge mode.

Table 5.3: Summary statistics collected for all Cuffmerge results and UCSC hg19 annotation.

Method	hg19	merged	merged_opt_g	merged_opt_s	merged_opt_gs
#genes	43,925	181,136	122,668	83,822	56,153
#isoforms	376,482	252,015	383,844	143,569	311,363
#isoforms per gene	8.57	1.39	3.13	1.71	5.54
Method		rabt.merged	rabt.merged_opt_g	rabt.merged_opt_s	rabt.merged_opt_gs
#genes		238,520	118,092	93,212	51,666
#isoforms		258,934	384,934	106,818	312,682
#isoforms per gene		1.09	3.26	1.15	5.05

To discover more of the pairwise consistency among all 8 merged assemblies, a distance matrix is computed. Similar with previous study, the similarity score between two modes i and j(could also be hg19 reference) is defined as:

$$\frac{\text{\#shared isoforms}}{\text{\#union isoforms from i and j}} \times 100\%$$

. The heatmap showed in Figure 5.9 is plotted according to this distance matrix. We see that although the statistics are quite comparable, actually very little similarity exists between Cufflinks without annotation and RABT assembly. But within each assembly mode group, the Cuffmerge results with reference transcripts are approximately consistent with each other: a similarity score around 0.8 can be observed from heatmap of option "-g" and "-gs".

In summary, great ambiguity rises when merging transcripts from large number of samples. Different options gave very different results, and the effect of each option is difficult to predict.

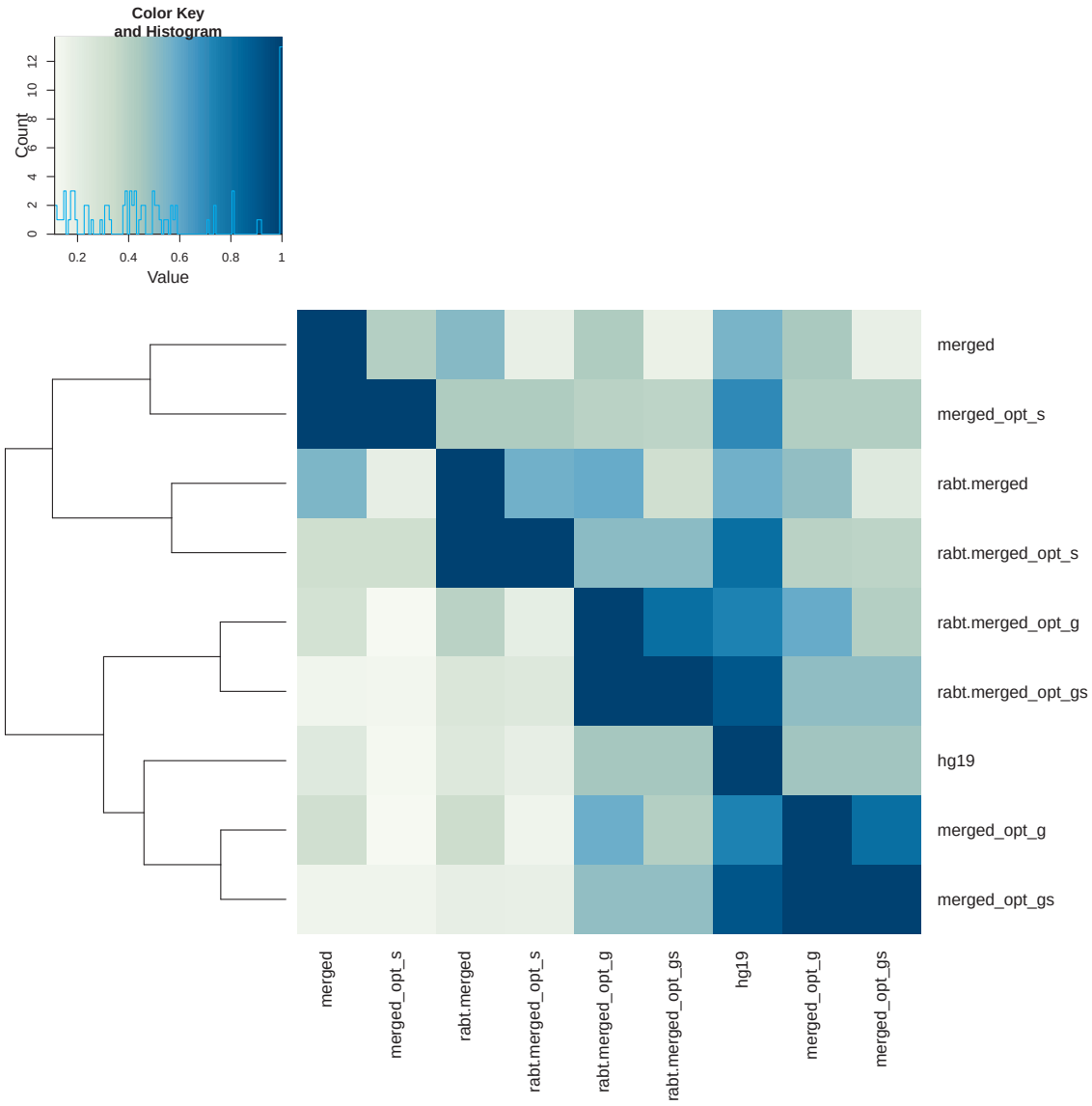


Figure 5.9: Heatmap comparing similarities among Cuffmerge results and UCSC hg19 annotation. Darker color means higher similarity.

5.2.3 Summary

In this section, we have investigated Cufflinks with 2 assembly modes and Cuffmerge with 4 assistance options. The comparison experiments have demonstrated that different modes of Cufflinks and Cuffmerge will yield very different results. The inconsistency among diverse Cufflinks+Cuffmerge pipelines would introduce great ambi-

guity and difficulty for downstream analysis, such as Cuffdiff which takes input of the merged transcript set for differential transcription analysis. Moreover, although the assemblies of each individual sample look very much alike, uncertainty of transcript reconstruction can be largely amplified when one tries to acquire the union of the entire assembly result. The merged set produced by Cuffmerge has considerable discrepancy with one sample.

Therefore, in general, although Cufflinks+Cuffmerge is a powerful pipeline that can be extended to the application on large-scale datasets, the complexity of the data and the uncertainty of the program could make the analysis intractable.

5.3 An *ab initio* method for the detection and visualization of differential transcription on large-scale dataset

Despite the shortages mentioned above, reconstruction and quantification of full-length transcripts are challenging problems due to the inadequate read length and sampling depth. The *unidentifiability* problem exists where no unique solution may exist for quantifying genes with multiple isoforms.

To circumvent all these problems of existing pipelines, we have developed a pipeline that performs comparative transcriptome analysis across hundreds of RNA-seq samples through a unified splice graph. This pipeline provides an *ab initio* method for the detection and visualization of differential transcription without the knowledge of transcript and/or gene annotations (Figure 5.2a). Our method describes a scalable approach to perform differential splicing analysis on large-scale (TBs and up) data. The input BAM files are first pre-processed locally to allow distributed data

storage. The processed expression information, in much reduced size and complexity, is collected from every sample and pooled to initiate the DiffSplice algorithm [Hu et al., 2012]. All samples are handled together in the analysis, as enabled by the unified expression-weighted splice graph (ESG). This could benefit difficult tasks such as junction filtering and transcript reconstruction by leveraging overall statistics of the whole data set. In comparison, the TCGA default RSEM pipeline Figure 5.2b and Cufflinks pipeline Figure 5.2c both handles each sample separately. The RSEM pipeline relies on a fixed transcriptome annotation and will discard novel isoforms. The Cufflinks pipeline, though performing quantification procedure on a consensus transcript set merged over all samples, essentially reconstructs transcripts for each sample individually.

Our work addresses challenges of large scale RNA-seq data analysis with the following improvements: (1) computational efficiency capable of detecting novel transcription variants and scalable with increasing sampling depth and rocketing number of samples, (2) accurate splice graph construction through joint analysis of all samples: The splice graph catalogues the set of exonic structures as well as high confidence alternative splicing events. We propose a junction classification method where putative junctions are classified based on their collective appearances from all samples. This allows us to remove false positive junctions with insufficient or inconsistent occurrence; (3) accurate differential splicing identification at the alternative splicing level: The splice graph provides a central reference for the mega-transcriptome. Each sample becomes an instance of the splice graph parameterized by its own expression at the splice graph features. To avoid the uncertainty of full length transcript inference,

the comparison between samples is done in the splice graph level by examining the difference of expression at the alternative splicing sites.

5.3.1 Method

To efficiently handle hundreds of samples that are terabytes in size, we have created a pre-processing step to condense the raw alignments (BAM files) into a much succinct format that is both effective and efficient for downstream analysis or reanalysis. Evey read alignment is split 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, are merged into one entry, associated with the count of its total presence. This compact count table requires only 15% of the size of the original data and allows fast reanalysis under one day for over 800 samples, yet maintains any necessary expression information for calculating the read distribution of the corresponding sample. Because each sample is processed individually, this procedure can be easily parallelized and accommodated to a distributed setup where the amount of data to be transferred and actively kept needs to be minimized. For the same reason, upon increase of data set, only the newly added samples need to be processed to facilitate a fast reanalysis.

Instead of independently reconstructing the transcriptome for every individual sample, our pipeline pools information from all samples first and establishes a unified structure of transcription for the entire data set. Splice junctions are filtered based on their overall occurrence – a putative junction is considered in the analysis for all samples even it has been seen, with adequate evidence, only in a small portion of the

samples. Pooled read coverage information averaged over all samples is examined, suggesting transcription start/end sites by recognizing positions with sharp coverage change. Together with the splice sites of the junctions, these sites are integrated to determine the precise exon boundaries in the genome. A unified splice graph is hence obtained by adding connectivity (edges) defined by splice junctions and retained introns between exons (vertices). An algorithm that utilizes the wavelet transformation has been developed as well to assist the removal of intergenic noises. The unified graph is lastly augmented into an Expression-weighted Splice Graph (ESG) with a weight vector to keep track of the observed read coverage of every exon and junction in each sample. Therefore, the ESG provides a central reference for the transcriptome in all samples, and each sample becomes an instance parameterized by its own expression at the graph components. With this strategy, the reconstruction algorithm needs to be applied for only once, and any two samples can be compared on a same structure. More importantly, for transcripts that are lowly expressed in some samples, they can still be recognized when other samples are leveraged where they are more abundant. This could benefit the completeness of the transcript reconstruction and the accuracy of the differential analysis.

With the ESG available, all possible ways that the exonic sequences of a gene can get transcribed have been depicted. Every graph path representing a possible 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. However, directly retrieving full-length transcripts often typically rely on heuristics [Trapnell et al., 2010b, Li et al., 2011b] to enforce

shrinkage on the resulting transcript set. Alternatively, we break down the complexity of the ESG iteratively into single-entry single-exit closed subgraphs, the alternative splicing modules (ASMs) [Hu et al., 2012]. This decomposition approach is especially advantageous in large-scale analysis in which transcript-based models are quickly overwhelmed by many heterogeneous samples. For every sample, the abundance of each alternative splicing path is inferred based on the expression of the exons and junctions within and surrounding the corresponding ASM.

The resulting profiles of splicing isoforms in each sample enable the comparison of the transcription patterns between subsets of samples, such as normal versus tumor and across tumor subtypes. In addition to the two-group comparison statistic, we have defined a multi-group test statistic for differential transcription of each ASM on the basis of Jensen-Shannon divergence. Analogous to the F-statistic, this test statistic tests the null hypothesis in which all groups have the same splicing profile on the ASM. A permutation test is then carried out to derive the null distribution of the test statistic for all ASMs across the whole genome. The deviation of the observed value of every ASM from the null distribution is then used as the evidence of differential transcription. Significant differences are selected according to the estimated false discovery rate (FDR) [Hu et al., 2012]. For significant alterations, pairwise differences are further evaluated and tested between subtype groups to search for alternative splicing variants that have consistent divergence. This procedure will highlight differential splicing events, the alternative splicing events that exhibit significantly different splicing ratios between different samples or between different subtypes.

Preprocessing. The goal of the preprocessing step is to condense the raw align-

ment files into a much succinct format that is both effective and efficient for downstream analysis or reanalysis. 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 a 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. The splitting and counting job for each sample can be performed in a distributed fashion, then the count tables are combined for the entire dataset using a linear-time merge sort-like algorithm. This strategy can minimize the data transfer and storage for analysis, and can easily accommodate new samples.

Construct the unified ESG. This pipeline utilizes a joint analysis model that summarizes all samples with a single graph, which can leverage information from all samples. 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.

Alternative splicing events discovery. The alternative exonic events can be identified through the decomposition of the ESG into ASMs (alternative splicing modules). 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. The decomposition of an ESG follows a 3-step process [Hu et al., 2012] which allows an iterative identification of all ASMs in the gene. Step 1: calculate the immediate pre/post dominators of every vertex of the ESG. Step 2: discover ASM whose entry or exist are vertices with out-degree or in-degree are more than 1. Step 3: discover nested ASMs where we iteratively identify nested ASMs within existing ones until no new ASMs can be found in Step 2.

Normalization across samples. The estimated expression of an alternative transcription path is determined by both its proportion in the transcriptome of the sample and the sampling depth of the RNA-seq experiment for this particular sample. Looking at the absolute expression levels of the alternative paths in an ASM, first the estimated expression levels should be adjusted in order to account for the variation of sampling depth among samples in the data set. We normalize the RNA-seq dataset by normalizing the total number of reads sequenced in each sample, which measures the sampling depth of a sample. 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.

Statistical test of differential splicing between multiple sample groups. Tumors have traditionally been classified into groups according to origins and subtypes. The different patients and subtypes may reflect different mixtures of cells and possibly different cancer mechanisms. Here we focus on the question whether there are consistent differences in isoform utilization between groups.

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,

$$\bar{\mathbf{q}}_\cdot = \frac{\sum_{j=1}^{k} n_j \cdot \bar{\mathbf{q}}_j}{\sum_{j=1}^{k} n_j}. \tag{5.1}$$

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

$$x_\Delta = \frac{\sum_{j=1}^{k} n_j \cdot JSD(\bar{\mathbf{q}}_j || \bar{\mathbf{q}}_\cdot)}{k-1} \tag{5.2}$$

and

$$s_\Delta = \frac{\sum_{j=1}^{k} \sum_{h=1}^{n_j} JSD(\mathbf{q}_j^h || \bar{\mathbf{q}}_j)}{\sum_{j=1}^{k} n_j - k}. \tag{5.3}$$

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

$$d_\Delta = \frac{x_\Delta}{s_\Delta + \sigma_\Delta}. \tag{5.4}$$

5.3.2 Experiment results

Landscape of alternative splicings in BRCA data. We have analyzed 819 RNA-seq samples from the TCGA breast cancer study (BRCA), including 728 tumor samples and 91 normal samples. The tumor samples are from five molecular subtypes of breast cancer including Luminal A, Luminal B, Basal-like, HEr2-enriched and Normal-like, and the normal samples are from two molecular subtypes Normal and Luminal A. The different patients and subtypes may reflect different mixtures of cells and possibly different cancer mechanisms. The original read alignments were first filtered to remove unannotated splice junctions found present (at least 10 spanning reads) in less than 5% of the data set (40 samples). A total of 237,823 junctions were kept out of the 1,153,635 raw junctions.

Determination of alternative splicing events.

It has been demonstrated that as many as 95% of all multi-exon genes are alternatively spliced during cell development, differentiation and diseases [Pan et al., 2008b, Wang et al., 2008b]. The mechanisms of the inclusion or exclusion of exons not only contributes to the biodiversity of proteins but is also considered as the implication of human genetic disorders, especially the development of cancer. A procedure has been designed to automatically categorize ASMs. All pairs of alternative paths in an ASM are organized into a priority queue [Cormen et al., 2001], with the most

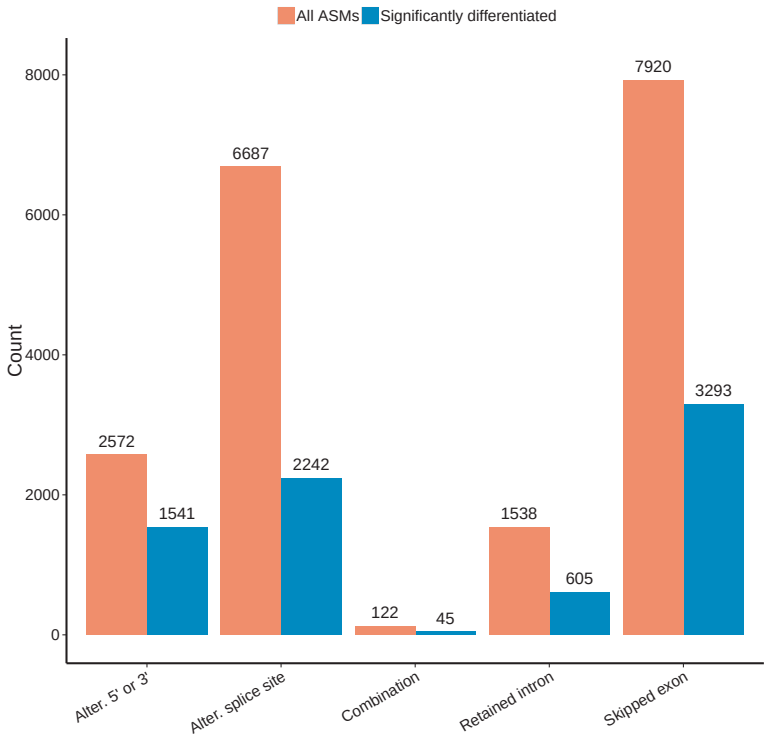


Figure 5.10: Alternative splicing events category plot on the entire TCGA breast cancer dataset.

abundant pair (sum of the abundance of the two paths in the pair) always examined first. The algorithm tries to match the pair, in the order of the queue, to one of the pre-defined basic alternative splicing models: skipped exon, mutually exclusive, retained intron, alternative splice site, alternative 5 or 3 end. The process will terminate when a match for a pair is found. If no basic model could fit, we describe the pair as "combined model".

In practice, quite many ASMs manifest a combination of multiple basic models, such as skipped exon and alternative splice site. Therefore, the priority queue is designed to suggest the dominant pattern that explains the majority of the expression within the specified genomic region. A total of 15,721 ASMs were discovered

and 7,309 of them involved novel junctions. Though many (¿5,000) events exhibited complex splicing models with more than two alternative isoforms, we were able to categorize the pattern with dominant expression and, consistent with previous findings [Eswaran et al., 2013], observed that exon skipping was the most common pattern with 7,030 occurrences, followed by 4,990 alternative splice sites, 2,907 alternative 5/3 transcription sites and 685 retained introns (Figure 5.10). Meanwhile, approximate 1/3 to 1/2 of ASMs are significantly differentiated for each category. This result illustrates that there is no bias towards any specific ASM category in the differential analysis.

Inter- and intra-subtype transcription patterns.

Controlled by an $FDR < 0.01$, the pipeline reported 7,262 differentially spliced loci from 5,442 genes, including 3,626 novel junctions. Focusing on the tumor samples only, the pipeline reported 4,293 differentially spliced loci from 3,510 genes under the same FDR, including 2,180 novel junctions. The heatmap in Figure 5.11 plots the expression of the 50 most differentially spliced genes, represented by their most variate alternative path. The normal samples generally exhibit large contrast from tumor samples, but also form two clusters. Partial normal samples, together with normal Luminal A samples, constitute a relatively distinct cohort from rest samples (leftmost columns). The remaining normal samples show similarity with a group of Luminal A samples. The Basal samples and the Luminal samples show large contrast. The majority of HEr2 samples exhibit concordance, whereas Luminal A and Luminal B samples tend to be more heterogeneous.

Subtype-specific regulation of alternative isoforms. The expression of

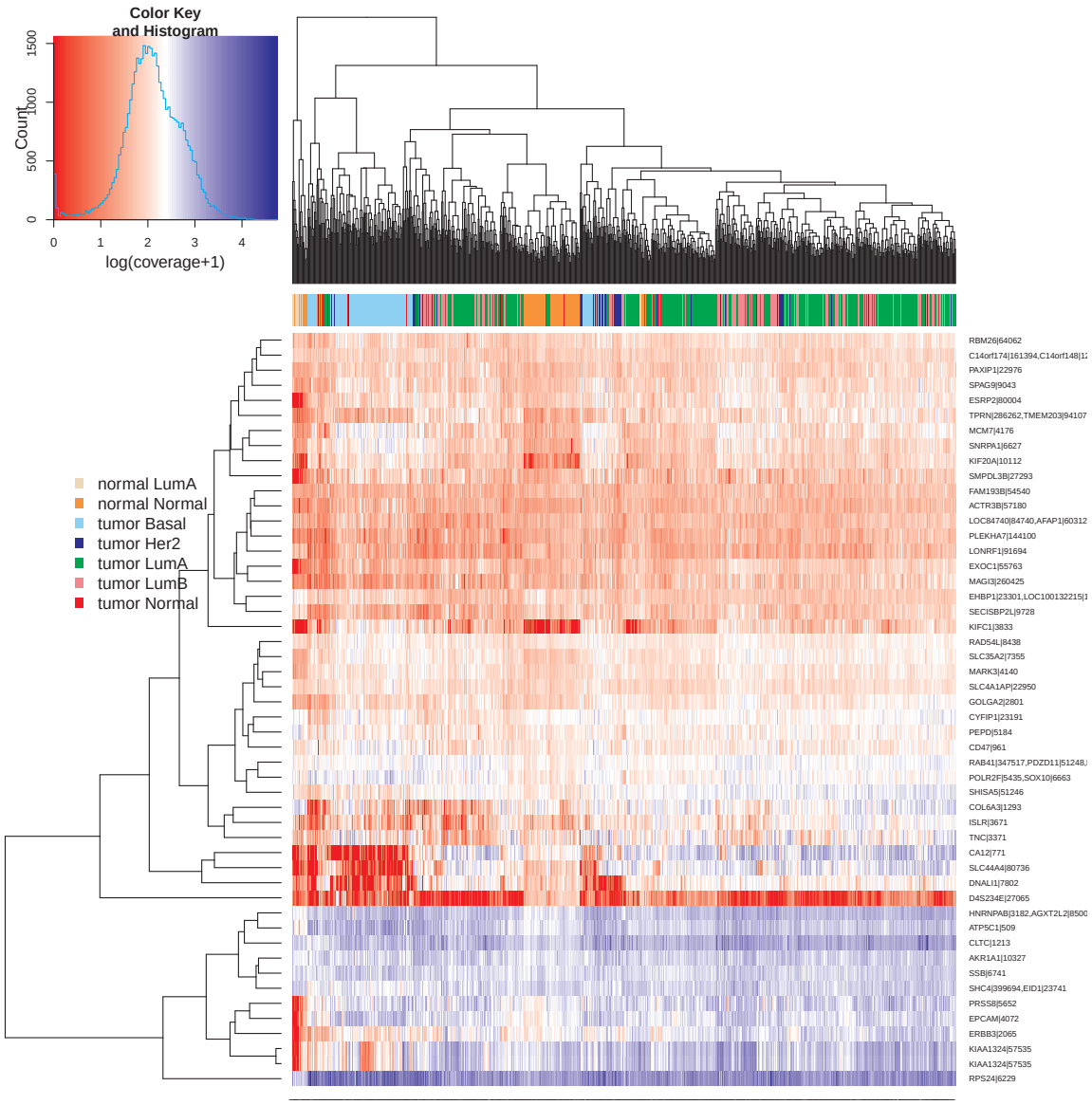


Figure 5.11: Heatmap of top 50 most differentially transcribed ASMs (represented by most divergent path) on the entire TCGA breast cancer dataset. The corresponding gene symbols are labeled on the right.

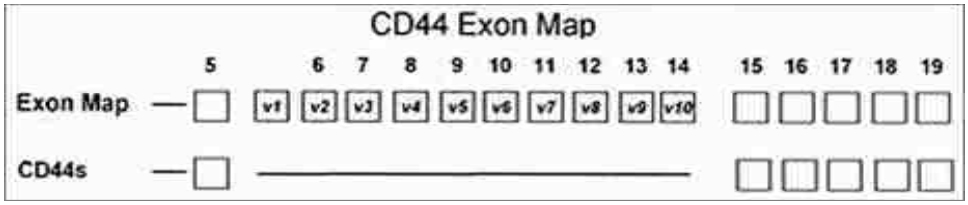


Figure 5.12: Exon map of gene CD44.

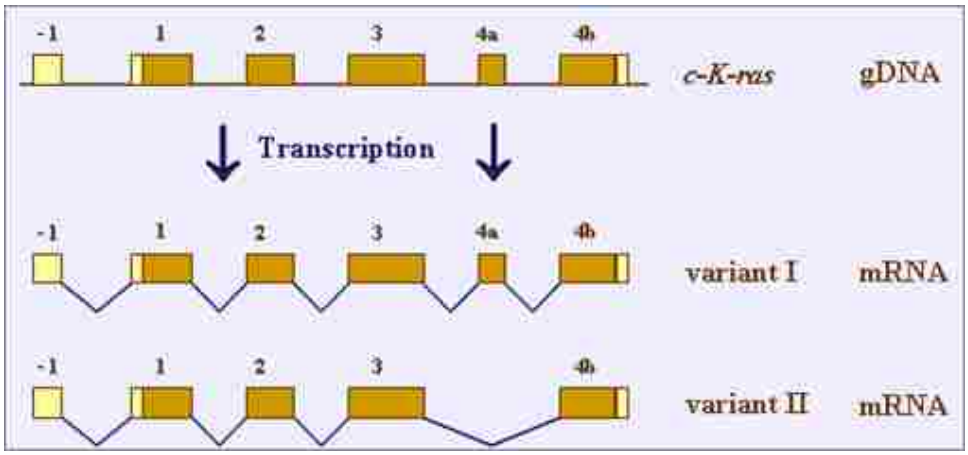


Figure 5.13: Two variants of gene KRAS.

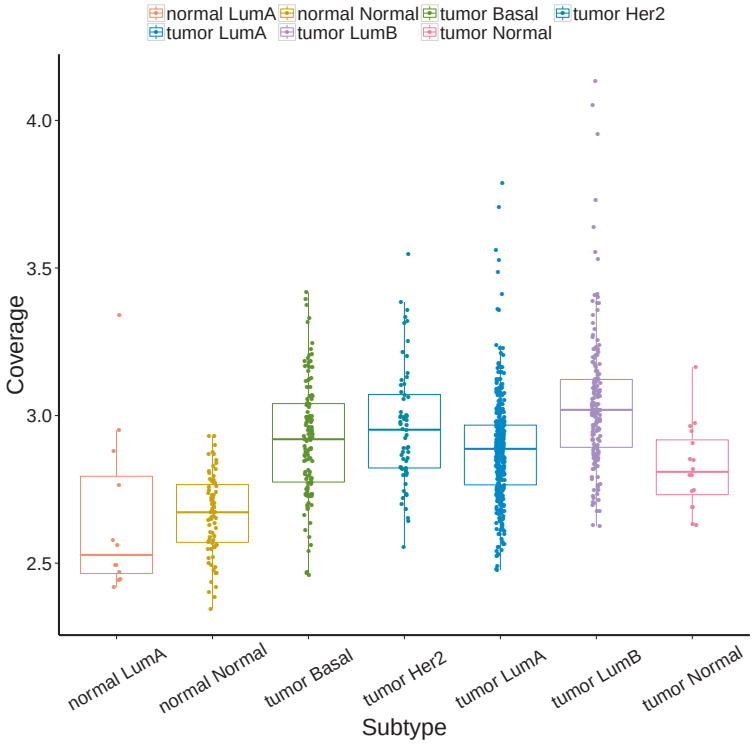


Figure 5.14: log 10 coverage of gene ErbB3.

CD44 isoforms have been suggested as markers of epithelial cancer stem cells. Its standard isoform, CD44s, skips all variant exons between exon 5 and 15 (Figure 5.12). High expression of CDD44s has been reported to be associated with strong HEr2 staining and a subgroup of basal-like tumors. Alternative isoforms have been an-

notated by retaining different sets of variant exons. In particular, previous study suggested the correlation of CD44v2-v10, the isoform retaining all variant exons, with low proliferation and Luminal A subtype. In our analysis (Figure 5.15), we confirmed higher percentage of CD44v2-v10 in Luminal A group ($P = 9.854e - 10$) and, as a comparison, an increased abundance of CD44s in HEr2 and Basal groups ($P = 8.183e - 14$).

Gene KRAS is a frequently mutated oncogene related to many types of cancer and KRAS mutations are often associated with a poor overall survival. Two variants are formed by an exon-skipping event and Variant I is known as a negative regulator of mutant KRAS alleles (Figure 5.16). In KRAS, we detected the exon skipping event that constitute the two known variants of this gene. The retaining path that keeps exon 4a corresponds to variant I, a variant known as a negative regulator of mutant KRAS alleles. We found over-expression of variant I in HEr2 and Luminal samples ($P < 2.2e - 16$), but not Basal and Normal-like tumors. This pattern may be associated with different mutation rates in breast cancer subtypes (Figure 5.16).

Potential modulator involving novel splice variant. The comprehensive analysis on cancer transcriptomes has also highlighted several novel alternative splicing events with subtype-specific transcription. An unannotated mutual exclusive event in gene CYFIP1 exhibited isoform switching in Luminal A and Luminal B as compared to Basal-like and HER2-enriched. Figure 5.17(a) displays the structure of the novel mutual exclusive event in one of Luminal B tumor sample. A unique exon (the second one) forms one alternative path that is missing from tumor Normal sample. This path is also shown as path $p1$ in Figure 5.17(b). We can see clearly

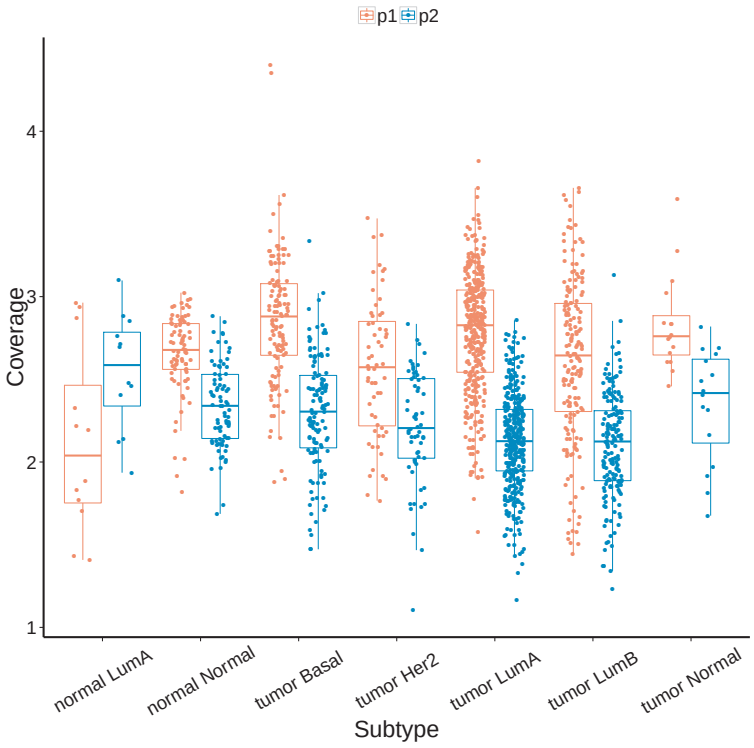
that $p1$ is more abundant in Luminal A and Luminal B samples than all the other samples.

Differential gene expression. In addition to detecting transcription level differences, The DiffSplice-based pipeline also estimates read coverage of every gene. Unlike gene level methods which directly counts the number of reads falling in one gene or averages the read coverage over the entire gene, our method propagates abundance information of splicing variants to the estimation of gene expression and therefore is unbiased toward different transcript length. For example, we detected up-regulation of gene ErbB3 in all tumor subtypes as compared to normal groups. In line with the previously reported correlation between ErbB3 expression and luminal breast cancer growth (ErbB3 downregulation enhances luminal breast tumor response to antiestrogens), the fold changes of Luminal B and Luminal A groups were 2.80 and 1.75, respectively as compared to double normal group. Moreover, we found over-expression of ErbB3 in Basal (fold change=1.89) and HEr2 (fold change=2.13) groups, suggesting its correlation with all breast cancer subtypes (Figure 5.14).

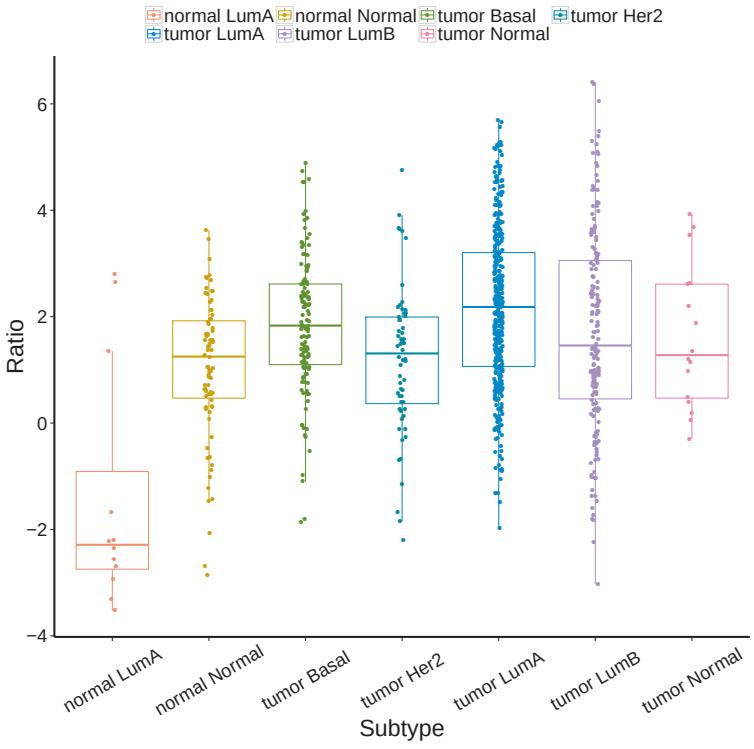
5.4 Discussion

With the rapid development of sequencing technology, more and more samples are being sequenced by researchers seeking for a comprehensive understanding of cell development or diseases. But how do we extract salient information from the massive dataset remains a challenging problem, for example, can we find biomarkers from hundreds of cancer transcriptomes? In this chapter, we review the problems with analyzing the large-scale RNA-seq dataset and systematically investigate the existing

pipelines using TCGA breast cancer dataset which consists of 819 RNA-seq samples. Given the limitations of the previous methods, we have developed a comparative model aiming at a unified analysis of the entire dataset. To our best knowledge, this is the first approach dedicated to large scale transcriptome analysis. Its contribution is three-fold. First, it's highly scalable. By preprocessing the high volume data first, it greatly condenses the raw data into a much succinct format which is supper effective and efficient for downstream analysis or reanalysis. Moreover, this process is easy to extend to distributed systems as well as deploy on cloud platform. Second, the unified framework is more accurate. It makes use of the power of full dataset which eliminates the biases introduced in the filtering step when merging all sample results together. Last but not least, it has a procedure automatically detect the alternative splicing category of the potential splicing signatures which manifests a visual idea of mechanisms of alternative exon selection.

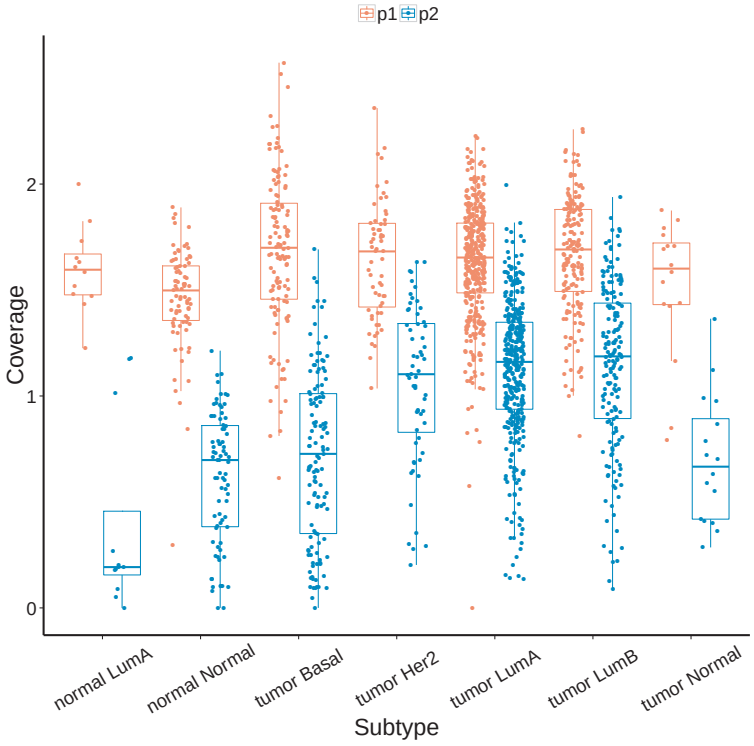


(a)

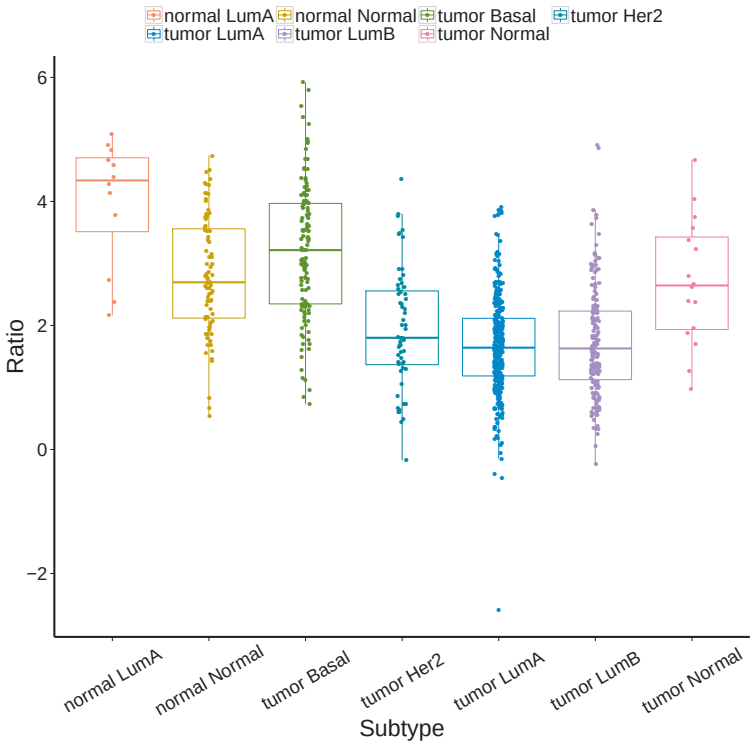


(b)

Figure 5.15: Gene CD44. (a) Boxplot showing the ASM path abundance distribution of all samples grouped by the subtype: CD44v2-v10 (orange) and CD44s (blue). (b) log2 expression ratio of CD44v2-v10 / CD44s.

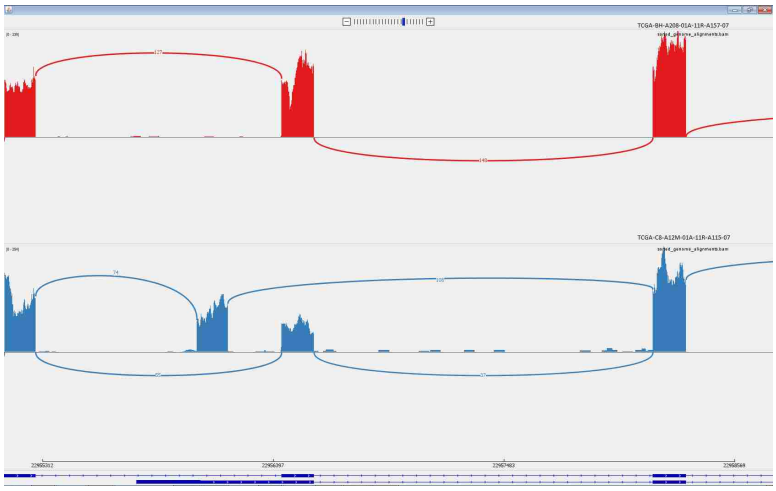


(a)

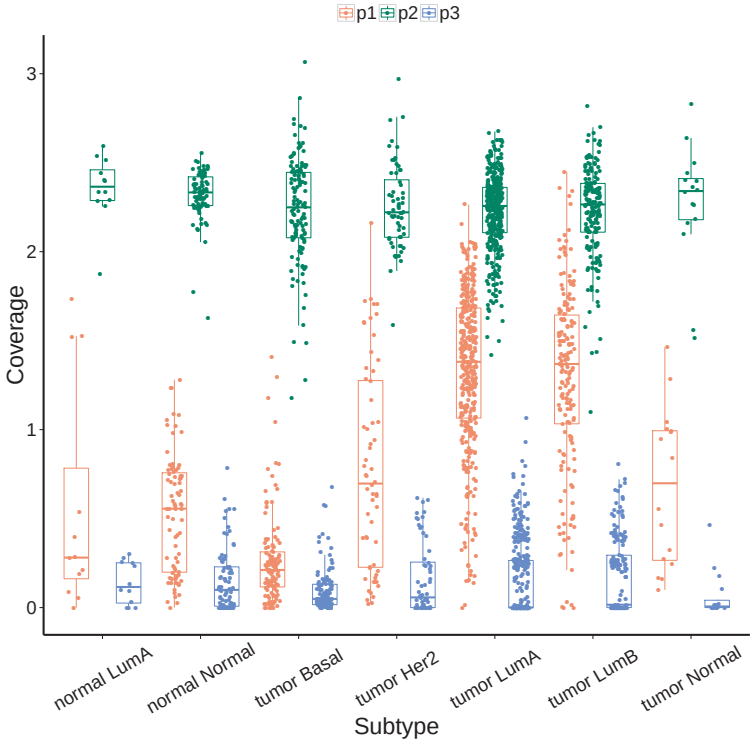


(b)

Figure 5.16: Gene KRAS. (a) Boxplot showing the ASM path abundance distribution of all samples grouped by the subtype: variant II (orange) and variant I (blue). (b) log2 expression ratio of variant II / variant I.



(a)



(b)

Figure 5.17: Gene CYFIP1. (a) Sashimi plot of the mutual exclusive event structure in gene CYFIP1. The upper plot shows a sample from tumor Normal and the bottom plot shows a sample from tumor Luminal B. (b) Boxplot showing the ASM path abundance distribution of all samples grouped by the subtype.

Chapter 6 Conclusion

This dissertation has presented a series of methods for a comprehensive analysis of mRNA transcriptome using RNA-seq data, including accurate abundance estimation of the annotated transcripts, simultaneous reconstruction and quantification of the novel genes and isoforms, and unified analysis on large-scale datasets. With these approaches, we are able to study the steady state of the existing mRNA transcript database as well as detect genes or isoforms that are not previously catalogued. Most importantly, an *ab initio* pipeline has been developed which takes advantage of the advancement of sequencing technology and performs transcriptome analysis both efficiently and effectively on massive RNA-seq data.

Precise transcript quantification determines the steady state levels of alternative transcripts within a sample, enabling the detection of differences in the expression of alternative transcripts under different conditions. Its application in detecting biomarkers between diseased and normal tissues can greatly impact biomedical research. The challenges of solving this problem reside in three aspects: first, short reads often do not uniquely identify the transcript isoforms from which they were sampled, leading to a unidentifiable model of transcript quantification; second, RNA-seq reads sampled from the transcriptome exhibit unknown position-specific and sequence-specific biases; lastly, not all of reference transcripts are likely to be significantly expressed in a given tissue type or condition. The first method has been developed aiming at resolving all these difficulties using a generalized linear model.

A novel feature named *MultiSplice* which summarizes reads spanning multiple splice junctions was incorporated as additional response variables to ameliorate identifiability. Bias parameters were also embedded into the coefficients of the linear model to adjust for various sampling biases. LASSO was further adopted for solving the linear system in order to infer an accurate set of dominantly expressed transcripts.

Simply studying the existing reference transcript database is not enough. Novel mRNA transcripts are often observed in various transcriptome, especially in diseased samples. For example, aberrant alternative splicing in cancer cells may lead to uncatalogued isoforms that could be potential biomarkers. To fully enable the translation from raw sequencing data to clinical insights, the core part of this dissertation focus on an *ab initio* framework to reconstruct the identities and quantities of the full-length gene isoforms residing in original cells, through inference based on the sequenced partial, noisy samples. We have developed a novel algorithm named *Astroid* that simultaneously infers isoforms together with their abundance, by assembling all observed reads into a set of chains that maximizes their joint probability. The observed data is modeled as a flow network, with every read being a vertex. Two reads are connected in the graph if they can potentially form a fraction of an isoform. The edges are further weighted by the likelihood of the linkage, evaluated according to the distribution of the distance. We solve for a best set of isoforms by finding a maximum likelihood set of non-overlapping source-to-sink paths, chains, which thread through all vertices. This maximization problem is then converted into a minimum-cost flow problem on the flow network. Lastly, we introduce a set of rules that clusters homogeneous vertices and edges and compresses the flow network. A compression parameter

is defined to leverage the time and space complexity required by the flow network and the model accuracy. Simulation studies and experimental benchmark on MAQC data set have demonstrated significantly improved sensitivity and specificity of Astroid in gene isoform reconstruction, as compared to 4 other state-of-the-art approaches. The unique design of Astroid that models a read-level network also enabled high accuracy in isoform abundance estimation. Furthermore, we have demonstrated the efficacy of the proposed compression method, which can successfully boost the speed and reduce memory usage by several magnitudes, while maintaining a comparable performance.

Empowered by the advancement of sequencing technologies, the role of sequencing data in any biomedical research/application is becoming more and more prominent. How to dig out treasures from the "big RNA-seq data" generated from a large number of samples is quite challenging. Besides effectiveness, scalability are now more emphasized in the evaluation of computational approaches. A comprehensive pipeline which extends our previous work *DiffSplice* is presented dedicated to large-scale transcriptome analysis. It has several contributions. First, since the complexity of data my increase exponentially with the sample size, the existing pipelines usually rely on known gene/isoform database by ignoring the possibility of uncatalogued transcripts in each sample. Our method, on the other hand, doesn't require any prior knowledge of reference transcripts, allowing the detection of novel genes and isoforms effectively. Second, our model directly takes input of the RNA-seq read alignments from all samples and constructs the unified genome-wide expression-weighted splice graph (ESG) to summarize the expression and splicing information on the genome in the given dataset. Downstream analysis is then performed on the unified ESG. While

other approaches all adopt per sample analysis first and then merge the results where great noises arise given the discrepancy among different samples. Moreover, further filtering may also lead to loss of information. Last but not least, our method iteratively decomposes the ESG into a set of ASMs: the regions where alternative splicing happens. These regions can then be further tested for potential splicing signatures between different group of samples. This procedure avoids the reconstruction and quantification of full-length transcripts which itself as demonstrated before is quite difficult especially consider the scale of the data.

All software packages of the methods described in this dissertation are open-source, released and maintained at `http://www.netlab.uky.edu/p/bioinfo/` and freely available to the research community.

Nowadays, the sample size is continuously increasing. For example, there were 510 samples when the first systematic investigation of TCGA breast cancer project was published in 2012 [Network, 2012], but now we are looking at totally 819 samples and the number is still climbing. More samples promise more power but also exhibit more variance which poses much more difficulty of a both efficient and effective analysis. The core unit of the extend DiffSplice pipeline handling massive dataset simultaneously is the construction of unified ESG which summarizes information from the entire dataset, filters out false signals and depicts all exonic segments and splice junctions on the genome. This process is also the foundation for Astroid since read alignments are connected into transcript copies guided by the possible path on the splice graph. Therefore, how to build an accurate splice graph with minimum false signals is very important, especially for large-scale dataset. Besides basic filtering

of spurious based solely on read count support, more sophisticated strategies should be investigated. One feasible approach would be use some collective statistics to guide the splice junction filtering, such as distribution of support in all samples, together with annotated gene models and find a reasonable way to select unannotated junctions that are probably real. More specially, a predicative model could be established as a binary classifier. Moreover, some complex cases are ignored during the construction of ESG, such as reads derived from total RNA-seq (where intronic reads dominate the sequencing library due to random priming instead of oligo-dT enrichment) and also fusion genes / chimeric transcripts, where reads / mate pairs map to different genes due to chromosomal rearrangements. However, these aberrant reads usually represent abnormal transcriptome activities. For example, fusion genes plays an important role in tumorgenesis because they can produce much more active abnormal proteins than non-fusion genes. Therefore, we could integrate this kind of information like fusion and extend this pipeline for a more complete transcriptomic profile.

Bibliography

Ensembl Genome Browser. `http://useast.ensembl.org/index.html`. 33

ICGC. `https://icgc.org/`. 92

NCBI Reference Sequence (RefSeq). `http://www.ncbi.nlm.nih.gov/RefSeq`. 33

TCGA. `http://cancergenome.nih.gov/`. 92

J. U. Adams. Transcriptome: Connecting the genome to gene function. *Nature Education*, 1:1, 2008. 6

R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows: Theory, algorithms, and applications. *Prentice Hall*, 1993a. 72

R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows: theory, algorithms, and applications. *Prentice Hall*, 1993b. 73

D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum1, and A. Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biology*, 12:R18, 2011. 68

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

G. Bejerano. Algorithms for variable length markov chain modeling. *Bioinformatics*, 20:788–789, 2004. 40

I. e. Birol. De novo transcriptome assembly with abyss. *Bioinformatics*, 25:2872–2877, 2009. 24

R. Bohnert and G. R. rquant.web: a tool for rna-seq-based transcript quantitation. *Nucleic Acids Research*, 38(Suppl 2): W348–W351, 2010. 16, 28, 30, 39, 46, 47, 78

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. 114

T. Clark, C. Sugnet, and M. Ares. Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296:907–910, 2002. 7

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2001. 116

A. Cox. Novoalign. ELAND: Efficient large-scale alignment of nucleotide databases. Illumina, San Diego., 2007. 25

M.-A. Dillies, A. Rau, J. Aubert, and *et. al*. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. 30

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. 25

J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, 19(2):248264, 1972. 72

J. Eswaran, A. Horvath, S. Godbole, S. D. Reddy, P. Mudvari, K. Ohshiro, D. Cyanam, S. Nair, S. A. W. Fuqua, K. Polyak, L. D. Florea, and R. Kumar. Rna sequencing of cancer reveals novel splicing alterations. *Scientific reports*, 3:1689, 2013. doi: 10.1038/srep01689. 118

R. Fisher, L. Pusztai, and C. Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer*, 108:479485, 2013. doi: 10.1038/bjc.2012.581. 92

K. L. Fox-Walsh, Y. Dou, B. J. Lam, S. pin Hung, P. F. Baldi, and K. J. Herte. The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci.*, 102(45):16176–16181, 2005. 46

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, 2011a. 25

M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. *Nature Methods*, 8:469–477, 2011b. doi: 10.1038/nmeth.1613. 24, 76

A. V. Goldberg and R. E. Tarjan. Finding minimum-cost circulations by canceling negative cycles. *Journal of the ACM*, 33 (Issue 4):873–886, 1989. 73

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. 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. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29:644–652, 2011. doi: 10.1038/nature07509. 24, 62, 76

T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guig, and M. Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic Acids Research*, 41:20, 2012. 66

M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, R. Corbett, M. J. Tang, Y.-C. Hou, T. J. Pugh, G. Robertson, S. Chittaranjan, A. Ally, J. K. Asano, S. Y. Chan, H. I. Li, H. McDonald, K. Teague, Y. Zhao, T. Zeng, A. Delaney, M. Hirst, G. B. Morin, S. J. M. Jones, I. T. Tai, and M. A. Marra. Alternative expression analysis by rna sequencing. *Nature Methods*, 7(10):843–7, 2010. doi: 10.1038/nmeth.1503. 77, 86

M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, 28:503–510, 2010. doi: 10.1038/nbt.1633. 26, 27, 62, 76

S. Heber, M. Alekseyev, S.-H. Sze, H. Tang, and P. A. Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, 18 (suppl 1):S181–S188, 2002. 26, 68

C. Hercus. Novoalign. www.novocraft.com. 25

C. Hiley, E. C. de Bruin, N. McGranahan, and C. Swanton. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. *Genome Biology*, 15:453, 2014. doi: 10.1186/s13059-014-0453-8. 92

K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. vant Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. V. Waes, Z. Chen, E. A. Collisson, T. C. G. A. R. Network, C. C. Benzemai, C. M. Perouemail, and J. M. Stuartemai. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158: Issue 4, p929944, 2014. doi: http://dx.doi.org/10.1016/j.cell.2014.06.049. 93

R. A. Horn and C. R. Johnson. Matrix analysis. *Cambridge University Press*, 1990. 38

Y. Hu, K. Wang, X. He, D. Y. Chiang, J. F. Prins, and J. Liu. A probabilistic framework for aligning paired-end rna-seq data. *Bioinformatics*, 26:1950–1957, 2010. doi: 10.1093/bioinformatics/btq336. viii, 33, 45

Y. Hu, Y. Huang, Y. Du, C. Orellana, D. Singh, A. Johnson, A. Monroy, P.-F. Kuan, S. Hammond, L. Makowski, S. Randell, D. Chiang, D. Hayes, C. Jones, Y. Liu, J. Prins, and J. Liu. Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic Acids Research*, 41(2):e39, 2012. doi: 10.1093/nar/gks1026. 26, 68, 109, 112, 114

Y. Huang, Y. Hu, C. Jones, J. MacLeod, D. Chiang, Y. Liu, J. 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. 26, 28, 62, 75

James. Rna-seq and the problem with short transcripts. *CoreGenomics*, 2011. 30

H. Jiang and W. H. Wong. Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, 25:1026–1032, 2009. doi: 10.1093/bioinformatics/btp113. 28, 32

M. R. Junttila and F. J. de Sauvage. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501:346354, 2013. doi: 10.1038/nature12626. 92

P. Kapranov. From transcription start site to cell biology. *Genome Biology*, 10(4):217, 2009. doi: 10.1186/gb-2009-10-4-217. 69

W. J. Kent. Blat-the blast-like alignment tool. *Genome Research*, 12(4):656–664, 2002. doi: 10.1101/gr.229202. 77

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

I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+c)-biased genomes. *Nature Methods*, 6:291–295, 2009. 38

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. 6

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9:357–359, 2012. 25

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, 2009a. 24, 57

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, 2009b. 25

B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:323, 2011. doi: 10.1186/1471-2105-12-323. 28, 30, 45, 67, 72, 80, 93

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. 16, 46, 96

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

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, 2008. 25

J. Li, H. Jiang, and W. H. Wong. Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biology*, 11, 2010b. 16, 39, 68

J. J. Li, C.-R. Jiang, J. B. Brown, H. Huang, and P. J. Bickel. Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci. USA*, 108(50):19867–72, 2011a. 44

W. Li, J. Feng, and T. Jiang. Isolasso: A lasso regression approach to rna-seq based transcriptome assembly. *15th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 6577:167–188, 2011b. 15, 26, 27, 29, 32, 62, 76, 78, 111

G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res.*, 21:936–939, 2011. 25

S. Marguerat, B. T. Wilhelm, and J. Bahler. Next-generation sequencing: applications beyond genomes. *Biochemical Society Transactions*, 36(5):1091–1096, 2008. 10

G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, page 114117, 1965. xi, 91

T. C. G. A. Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:6170, 2012. doi: 10.1038/nature11412. 93, 130

T. C. G. A. R. Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497:6773, 2013a. doi: 10.1038/nature12113. 93

T. C. G. A. R. Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511:543550, 2013b. doi: 10.1038/nature13385. 93

T. C. G. A. R. Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45: 11131120, 2013. doi: 10.1038/ng.2764. 93

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. 28, 46

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. 39

J. B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78: 109–129, 1997. 73

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, 2008a. 6

Q. Pan, O. Shai, L. J. Lee, and et.al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40:1413 – 1415, 2008b. doi: 10.1038/ng.259. 14, 116

E. Phizicky, P. I. H. Bastiaens, H. Zhu, M. Snyder, and S. Fields. Protein analysis on a proteomic scale. *Nature*, 422:208–215, 2003. 5

K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37:D32–D36, 2008. doi: 10.1093/bioinformatics/bts260. 85

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 Research*, 38:e112, 2010. doi: 10.1093/nar/gkq041. 44

A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011. doi: 10.1186/gb-2011-12-3-r22. 16, 30, 34, 39, 40, 68, 86

M. F. Rogers, J. Thomas, A. S. Reddy, and A. Ben-Hur. Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data. *Genome Biology*, 13 (Issue 1):R4, 2012. doi: 10.1186/gb-2012-13-1-r4. 68

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. 25

S. Russell and P. Norvig. Artificial intelligence: A modern approach. page R22, 2003. 42

G. Russo, C. Zegar, and A. Giordano. Advantages and limitations of microarray technology in human cancer. *Oncogene*, 22: 64976507, 2003. 7

L. Shi, R. LH, J. WD, S. R, and et.al. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006. 77, 84

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. 8

D. Singh, C. F. Orellana, Y. Hu, C. D. Jones, Y. Liu, D. Y. Chiang, J. Liu, and J. F. Prins. Fdm: A graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics*, 27 (19):2633–2640, 2011. doi: doi:10.1093/bioinformatics/btr458. 55

S. Srivastava and L. Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Research*, 38:e112, 2010. doi: 10.1093/nar/gkq041. 39

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. 6

B. Tenchov, T. Yanev, M. Tihova, and R. Koynova. A probability concept about size distributions of sonicated lipid vesicles. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 816(Issue 1):122–30, 1985. 66

H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14 (2):178–192, 2013. doi: 10.1093/bib/bbs017. x, 63

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B.*, 58:267–288, 1996. 31, 79

C. Trapnell, L. Pachter, and S. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25:1105–1111, 2009a. 68

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9): 1105–1111, 2009b. 24, 25

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25:1105–1111, 2009c. xii, 98

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, 2010a. doi: 10.1038/nbt.1621. 7

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010b. doi: 10.1038/nbt.1621. 26, 27, 28, 29, 62, 66, 76, 96, 97, 111

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010c. doi: 10.1038/nbt.1621. 45, 49, 55

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl. Acad. Sci. USA*, 99(22):1403114036, 2002. 9

E. Turro, S.-Y. Su, A. Concalves, L. J. Coin, S. Richardson, and A. Lewin. Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biology*, 12:R13, 2011. 16, 68

V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995. 8

G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory Biosci.*, 131(4):281–5, 2012. doi: 10.1007/s12064-012-0162-3. 30, 66, 67

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, 2008a. 6, 7

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge1. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008b. doi: 10.1038/nature07509. 15, 116

G. Wang and T. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, 8:749–761, 2007. 7

K. Wang, D. Singh, Z. Zeng, Y. Huang, S. Coleman, G. Savich, X. He, P. Mieczkowski, S. Grimm, C. Perou, J. MacLeod, D. Chiang, J. Prins, and J. Liu. Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acids Research*, 38 (18):178, 2010a. viii, xii, 24, 33, 45, 57, 68, 98

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, 2010b. 25

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10: 57–63, 2009a. 38

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63, 2009b. 8, 9, 10

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. 25

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, 2011. 16

Z. Xia, J. Wen, C.-C. Chang, and X. Zhou. Nsmap: a method for spliced isoforms identification and quantification from rna-seq. *BMC Bioinformatics*, 12:162, 2011. doi: 10.1186/gb-2012-13-1-r4. 68

R. Yamashita, N. P. Sathira, A. Kanai, K. Tanimoto, T. Arauchi, Y. Tanaka, S. ichi Hashimoto, S. Sugano, K. Nakai, and Y. Suzuki. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Research*, 21(5):775789, 2011. doi: 10.1101/gr.110254.110. 69

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. 25

Vita

Name

Yan Huang

Education

M.Sc., Statistics — May, 2013
University of Kentucky — Lexington, KY, USA
B.Eng., Computer Science — July, 2009
University of Science and Technology of China — Hefei, China

Publications

1. **Yan Huang**, Yin Hu, and Jinze Liu. Piecing the Puzzle Together: a Revisit to Transcript Reconstruction Problem in RNA-seq. ***BMC bioinformatics***, 2014, 15(Suppl 9):S3.

2. **Yan Huang**, Yin Hu, and Jinze Liu. Piecing the Puzzle Together: a Revisit to Transcript Reconstruction Problem in RNA-seq. ***4th Annual RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-SEQ)***, March 31 - April 1, 2014, Pittsburgh, USA.

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. ***Journal of Computational Biology***, 2013, doi: 10.1089/cmb.2012.0230.

4. Yin Hu, **Yan Huang**, Ying Du, Christian F. Orellana, Darshan Singh, and *et.al.* DiffSplice: the Genome-Wide Detection of Differential Splicing Events with RNA-seq. ***Nucleic Acids Research***, 2012, doi: 10.1093/nar/gks1026.

5. **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)***, April 21 - April 24, 2012, Barcelona, Spain.

6. Kai Wang, Darshan Singh, Zheng Zeng, Stephen Coleman, **Yan Huang**, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. Accurate Mapping of RNA-seq Reads for Splice Junction Discovery. ***Nucleic Acids Research***, 2010, doi: 10.1093/nar/gkq622.

Conference Presentations

1. **Yan Huang**, Yin Hu, and Jinze Liu. Simultaneous transcript reconstruction and quantification using paired-end RNA-Seq reads. ***UT-ORNL-KBRIN Bioinformatics Summit 2014***, Cadiz, USA, April, 2014. (Oral presentation)

2. **Yan Huang**, Yin Hu, and Jinze Liu. Piecing the Puzzle Together: a Revisit to Transcript Reconstruction Problem in RNA-seq. ***4th Annual RECOMB***

Satellite Workshop on Massively Parallel Sequencing (RECOMB-SEQ), Pittsburgh, USA, April, 2014. (Oral presentation)

3. 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)

4. **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)

5. 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.

6. **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)

7. 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)

8. **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.

9. David Fardo, **Yan Huang**. Testing Gene-Environment Interactions in Family-Based Genetic Asso- ciation Studies: A Causal Inference Approach to Adjust for Ascertainment-induced Bias. ***UT-ORNL-KBRIN Bioinformatics Summit 2010***, Cadiz, USA, March, 2010.

Software

Astroid: simultaneous reconstruction and quanti

cation of gene isoforms using paired-end RNA-seq data.

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

MultiSplice: inferring the abundance of gene isoforms based on observed read coverage.

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

Awards

Kentucky Opportunity Fellowship	2014
Nominee of Dissertation Year Fellowship in University of Kentucky	2014
National Science Foundation (NSF) travel grant for RECOMB	2012
CRA-W Travel Grant	2010
Outstanding Student Scholarship	2006, 2007, 2008
Undergraduate Research Project Excellent Award (Top 1)	2008
Ranked 1st in Undergraduate Research Program (supported by Microsoft)	2007