GENOME WIDE COMPUTATIONAL ANALYSIS OF ALTERNATIVE SPLICING WITH THE PASTA PIPELINE

By

SHAOJUN TANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

To my wife, Wei Wang, who always encouraged me in the hardest days and provided me with constant support, and to my parents and supervisor for their help in pursuing my goals and dreams.

ACKNOWLEDGMENTS

In writing the acknowledgment, I would like to thank for all of the members of the committee, my colleagues in Genetics and Genomic graduate program, and my family, with the help from all of them who make the dissertation possible.

I would like most to thank my advisor professor Alberto Riva. I am able to choose the area of bioinformatics as my fields of study with supervise from him. He guided me into bioinformatics and nourished my interest on this field. He taught me programming skills and helped me overcome difficulties of various kinds. I am so fortunate to have such a great supervisor who sets the standard for my career.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Adenine
AS	Alternative splicing
BPS	Branch point sequence
C	Cytosine
DM1	Myotonic dystrophy 1
DNA	Deoxyribonucleic acid
EST	Expressed sequence tags
FN	False negative
FNR	False negative rate
FP	False positive
FPR	False positive rate
G	Guanine
GA	Illumina genome analyzer
GLM	Generalized linear model
GNU	GNU’s is not UNIX
GO	Gene ontology
KD	Kilo Dalton
LOG	Logarithm
MBNL2	Muscleblind-like protein 2
MBNL3	Muscleblind-like protein 3
mRNA	Message RNA
NMD	Nonsense mediated decay
NT	Nucleotide
ORF	Open reading frame

PCR	Polymerase chain reaction
PPT	Polypyrimidine tract
R	GNU statistical computing software
RAM	Random access memory
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
SMA	Spinal muscular atrophy
SNP	Single nucleotide polymorphism
SQL	Structured query language
SRA	Short Read Archie
T	Thymine
TN	True negative
TP	True positive
UTR	Un-translated region

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

GENOME WIDE COMPUTATIONAL ANALYSIS OF ALTERNATIVE SPLICING WITH THE PASTA PIPELINE

By

Shaojun Tang

December 2012

Chair: Alberto Riva
Cochair: William Bradley Barbazuk
Major: Genetics and Genomics

There is increasing evidence of the correlation between disrupted regulation of alternative splicing and disease. Computational analysis of alternative splicing using RNA-Seq technology is playing an important role in studying the spatial and temporal expression patterns of gene isoforms and analyzing the relationships between these alternative splicing patterns and disease phenotypes.

To have a better understanding of splicing regulations in comparison of normal and disease phenotypes, I conducted a genome wide computational analysis on alternative splicing to study the transcriptome composition and regulation of gene isoforms. Existing software packages cannot effectively predict splice junctions and identify genes with differentially expressed isoforms. Therefore, I proposed new algorithms to study splicing and alternative splicing by developing a software pipeline for splice junction detection, isoform reconstruction and gene expression estimation, using RNA-Seq datasets from a wide range of studies. In addition, I performed a detailed alternative splicing analysis using RNA-Seq datasets from *Mbnl3* and *Mbnl2* isoform knockout mice to find alternative splice junctions and identify differentially regulated

isoforms. Experimental results showed that the software pipeline can effectively predict splice junctions including minor splice junctions and identify genes with differentially expressed isoforms. The application of the software pipeline can provide a genome-wide characterization of alternative splicing and its regulation, and can be used to analyze changes in RNA isoform expression levels using RNA-Seq data.

CHAPTER 1
INTRODUCTION

This chapter provides background information on basic nucleic acid bases, splicing, alternative splicing and its association with disease. I start by providing basic concepts of genetic information in DNA sequences, followed by an introduction to splicing, alternative splicing and its regulation. Finally, I give a few examples of diseases that are associated with splicing.

Biological Concepts of Genetic Information

Living organisms carry genetic information in the form of nucleic acid sequences. The discovery of the double-helix structure of DNA in 1953 and subsequent studies helped scientists make great strides in our understanding the role of DNA in genetic disorders and diseases [1].

DNA sequences are composed of two complementary strands of nucleic acids. The four nucleobases are Guanine (G), Cytosine (C), Adenine (A) and Thymine (T). A nucleotide consists of a nucleobase, a five-carbon sugar and a phosphate group. Each nucleotide is linked to the next by ester bonds in the same strand, and each nucleotide is base paired to the complementary nucleotide in the other strand by hydrogen bonds as shown in Figure 1-1 [2]. Instead, RNA contains the nucleobase uracil (U) which corresponds to thymine (T) in DNA. RNA plays an important role in gene expression, cellular communication, and as template for protein synthesis.

The classic central dogma of molecular biology indicates that genetic information flows from DNA to RNA and finally to proteins as shown in Figure 1-2 [2, 3]. In addition, there are a large number of post transcriptional modifications taking place, such as RNA splicing, and post translational ones, such as phosphorylation of amino acids. Studies

on these regulatory mechanisms are essential for better understanding the flow of the genetic information and its regulation [4, 5].

Splicing and Alternative Splicing

Splicing

Splicing is the modification of nascent pre-mRNA transcribed from DNA sequences in generating mature mRNAs. Splicing, that was discovered in the 1970s by studies on adenoviruses [6], is controlled by the spliceosome, a complex consisting of five small nuclear RiboNucleoProteins (snRNPs) and a large group auxiliary proteins that interact with each other in catalyzing the splicing reactions [1, 7]. Pre-mRNA is cleaved at the 5' end of intron, and the removed intron sequence is attached to the branch site through base pairing to form a lariat structure [8]. In spliceosome assembly, the cleavage of intron is characterized by the presence of GU at the 5' splice site and AG at the 3' splice site [1, 9]. Therefore, these GU-AG (or GT-AG in DNA sequences) signals are termed *canonical splicing signals* and this splicing pathway is termed *canonical splicing*. There is another splicing pathway where one of the key snRNPs U2 is replaced by U12 in the splicesome composition which causes an alternative usage of splice signals. This is termed *minor splicing pathway* because it takes place much less frequently compared with canonical splicing in the major splicing pathway [1].

Splicing Signal

The selective usage of canonical or non-canonical splicing signals is critical to alternative splicing regulation [10]. It is estimated that more than 98% of exon junctions in eukaryotic organism utilize the GT-AG canonical splice signal [11]. However, studies also show that GC-AG splice site has a very strong splice donor consensus. The GC-AG splice signal is observed in human, nematode and Arabidopsis [12]. Replacement of

GC donor signal by GT signal has been associated with disrupted usages of certain exons in genes such as let-2 which is crucial for alternative splicing regulations [13]. Besides canonical signals and the GC-AG signals, other splice signals such as AT-AC, GT-TG, GT-CG or CT-AG are also observed [14].

Alternative Splicing

Alternative splicing is the process by which a single gene can generate multiple transcripts, and therefore different proteins, through the alternative use of exons, splice sites as shown in Figure 1-3 [1, 15]. As our knowledge of the structure and organization of genomes increases, Alternative splicing is being increasingly recognized as a fundamental process at the basis of the molecular diversity and complexity within the cell, of gene regulation, and of a number of critical biological processes ranging from development to disease [16]. Alterations of alternative splicing are linked to human diseases ranging from cancer to muscular dystrophies, from neurodegenerative diseases to obesity [15]. A better understanding of the mechanisms that regulate alternative splicing and of the relationships between alternative splicing and pathological states will provide new, important insights into these diseases, leading to advances in their diagnosis, and opening the way for the development of novel molecular therapies.

Alternative splicing is not only commonly observed in human, mouse and other higher organisms, but it is also a ubiquitous cellular process used by plants and even single-cell organisms such as yeast. Maize is one of the most widely investigated and agriculturally important plants because of its commercial value and its role as a major food source for very large populations in many countries [17]. Maize is also one of the largest plant genomes sequenced. Alternative splicing in maize shares many features with other organisms such as mouse, but there are also distinct features such as

prevailing presences of transposons and smaller gene sizes [18]. The B73 maize genome sequence was improved during a recent genome re-sequencing effort and over 32,000 genes were predicted [17]. This sequence assembly sets a new stage for further investigation of alternative splicing regulations for better understanding of the maize genome.

Regulation of Alternative Splicing

Alternative splicing regulations, such as alternative transcription start site selection, cassette exon usage and alternative polyadenylation, play an important role in generating the large diversity of higher eukaryotic transcriptome from a relatively small number of genes [19]. Different alternative isoforms produced by alternative splicing can have distinct structural features, functional properties and cellular localizations [20]. Alternative splicing is regulated by trans-acting components such as serine/arginine-rich proteins (SR proteins) and cis-acting sequences [21]. Splice site recognition characterized by splice signals directly contributes to the production of different alternative splicing junctions and isoforms. In addition, the different enrichments of intron or exon regulatory sequences such as Exonic Splicing Enhancer (ESE) and Exonic Splicing Silencer (ESS) play an important role in alternative splicing decisions [9]. Comparative analysis from human and mouse genomic sequences shows that selective pressures from evolutionary forces are other important factors regulating constitutive and alternative splicing [22].

RNA Quality Control by NMD

Alternative splicing plays an important role in mRNA quality control, including regulating spatial and temporal specific transcription. Alternative isoforms containing premature termination codons (PTC) can be formed as a result of alternative splicing

regulations. Nonsense- Mediated Decay (NMD) targeted degradation may be linked to the presence of PTC as a result of alternative splicing regulation [23, 24]. NMD pathway triggered by PTC is widely used for mRNA quality control to exclude deleterious gain-of-function or dominant-negative transcripts [23, 25]. The Exon-Junction Complex (EJC), a large protein complex that is assembled ~22 nucleotides upstream of the splice junction, is an important secondary signal that elicits the NMD pathway [26]. PTCs that are more than 55 nucleotides away from the downstream EJC are likely to induce NMD.

Alternative Splicing and Diseases

Disruption of mRNA splicing regulation has been associated with a large number of diseases as shown in Table 1-1 [15]. Typical causes of RNA related diseases include the use of cryptic alternative splice sites and the disruption of auxiliary splicing factors that can result in the changes of protein-RNA binding properties as shown in Figure 1-4 [27, 28]. Mutations in cis-acting and/or trans-acting factors, which are important regulatory elements for alternative splicing decisions, have been shown to be linked to a number of diseases such as cystic fibrosis [29]. Cis-acting factor mutations are associated with splicing code disruptions and trans-acting factors could lead to abnormal assembly and disrupted regulation of the splicing machinery. In addition, alternative splicing can affect gene expression levels and relative isoform expressions which in turn are associated with a number of diseases [28].In general, nucleotide mutations leading to disruption of mRNA splicing may underlie more than 15% of human disease [30, 31].

Diseases associated with disrupted MuscleBlind-Like Proteins (MBNL)

Myotonic Dystrophy (DM), one of the splicing-related diseases, is a complex neuromuscular disorder caused by CTG repeat expansions in Myotonic Dystrophy

Protein Kinase (MDPK) gene (Table 1-1) [32]. Manifestations of DM include heart block, ocular cataracts, nervous system dysfunction and muscle wasting. Previous studies showed that the functions of RNA-binding proteins, such as members of the MBNL family [33], are altered in this disease.

Mbnl3

Mbnl3 (also named *CHCR*, or *Cys3His-CCG1-Required*) is a member of the muscleblind-like family of *Cys3His* zinc finger proteins [34, 35]. Studies show that *Mbnl3* expression is primarily detected in placenta , myoblast and muscle precursor cells, but is also enriched in lung and spleen[36] *Mbnl3* is involved in regulating expression levels of the muscle transcription factor *Myod*. Microarray profiling experiments reveal that genes involved in muscle development and cell adhesion are negatively regulated by *Mbnl3* expression[34]. Myosin Heavy Chain (MyHC) differentiation is inhibited upon the expression of *Mbnl3* in C2C12 (Mouse myoblast) cells lines[36]. Therefore, instead of promoting muscle differentiation, *Mbnl3* expression in C2C12 cells prohibits muscle differentiation [34, 35].

Transcriptome composition analysis helps improve alternative splicing studies

Characterization of the transcriptome composition plays an important role in the study of spatial and temporal expression patterns of gene isoforms and proteins [37]. Differential inclusion of exons and alternative usage of splice sites causes more than 80% of human genes to produce multiple isoforms with distinct biological functions and composition [38]. Disruptions in splicing regulation are believed to be among the major reasons for neurological and neuromuscular diseases such as Spinal Muscular Atrophy (SMA) and Myotonic Dystrophy (DM) [39]. In particular, characterization of the mRNA

population using RNA Sequencing (RNA-Seq) can uncover important splicing regulation information in the study of pathological states.

Studying Transcriptome Composition and Splicing Regulation using RNA-Seq

RNA-Seq provides a robust approach to characterize the transcriptome and estimate isoform expression levels [40]. It enables us to discover new splice transcripts and identify genes with differentially expressed isoforms, which in turn leads to a better understanding of splicing regulations in disease phenotypes [41]. RNA-Seq offers several key advantages. First, it facilitates the analysis of transcriptome in less studied non-model organisms because it does not require the probe design step which is necessary in microarray technology. Second, it contains precise sequence information that can reveal sequence variants and polymorphisms. Third, novel splice junctions and transcript variants can be identified and quantified to provide unbiased alternative splicing analysis with complex disease phenotypes [42].

Existing Computational Pipelines for Alternative Splicing Analysis using RNA-Seq Dataset

The rapid application of RNA-Seq technology and large volume of DNA short read sequences require effective bioinformatics tools to identify the transcriptome composition and characterize splicing regulation. There are a number of available computational pipelines to analyze RNA-Seq datasets. Pipelines such as Bowtie provide a fast and memory efficient way in aligning RNA-Seq datasets to a reference genome [43]. The splice junction prediction tool Tophat is used to find exon junctions with canonical splice signals using the seed-and-extend algorithm [44]. Estimation of isoform expression levels and identification of genes with differentially expressed isoforms are summarized in pipelines such as MISO and Trans-ABySS [40, 45]. Overall,

these pipelines allow us to find alternative splicing transcripts and study splicing regulation using RNA-Seq datasets.

Development of an Innovation Software Pipeline for Splicing Analysis

In this work, I developed a novel computer pipeline, called PASTA (Patterned Alignments for Splicing and Transcriptome Analysis), to study alternative splicing regulation. This pipeline enables comprehensive alternative splicing analysis by providing splice junction prediction, *de novo* isoform reconstruction and identification of genes with differentially expressed isoforms. In comparison with existing tools, PASTA can effectively identify splice junctions, including splice junctions with minor splice signals, from RNA-Seq datasets. In addition, PASTA includes a novel feature for isoform reconstruction by utilizing codon usage preferences. Finally, I conducted a detailed alternative splicing analysis to identify genes with differentially expressed isoforms using data from case-control studies in disease phenotypes.

Table 1-1. Examples of diseases that are closely linked to alternative splicing.

Disease	Link to alternative splicing
Myotonic dystrophy (DM)-- DM 1	CUG expansion in the 3'UTR of DMPK results in the misregulation of the MBNL splicing factor and consequent missplicing of MBNL target pre-mRNAs
Spinal muscular atrophy (SMA)	Deletion/mutation of the SMN1 gene, and the loss of a splicing regulatory element in SMN2 results in insufficient levels of SMN, which is involved in snRNP biogenesis
Prader Willi syndrome	Loss of a splicing regulatory snoRNA that is complementary to a splicing silencer element implicated in regulating the alternative splicing of serotonin receptor 5-HT2cR transcripts

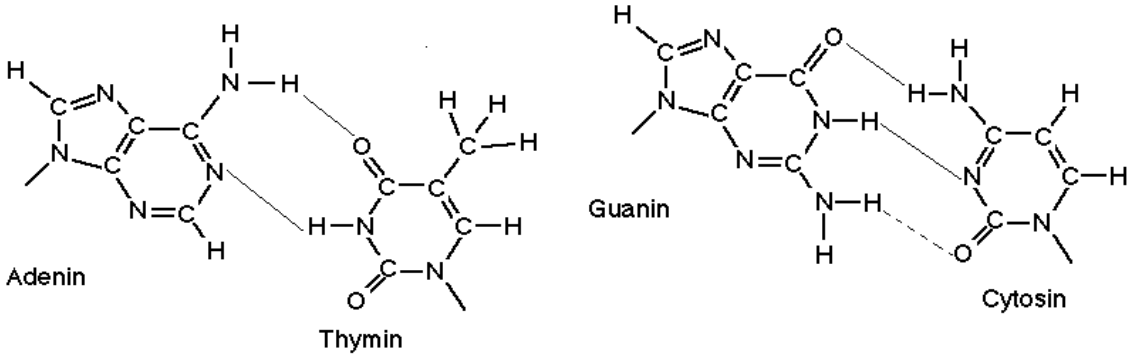


Figure 1-1. The basic nucleic acid bases.

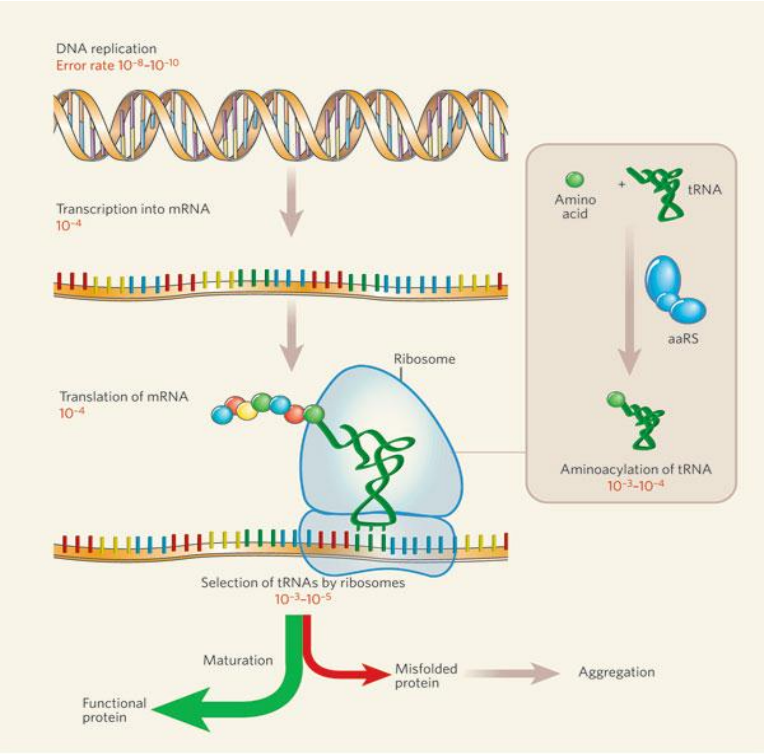


Figure 1-2. Information flow from DNA to RNA and to amino acid sequences.

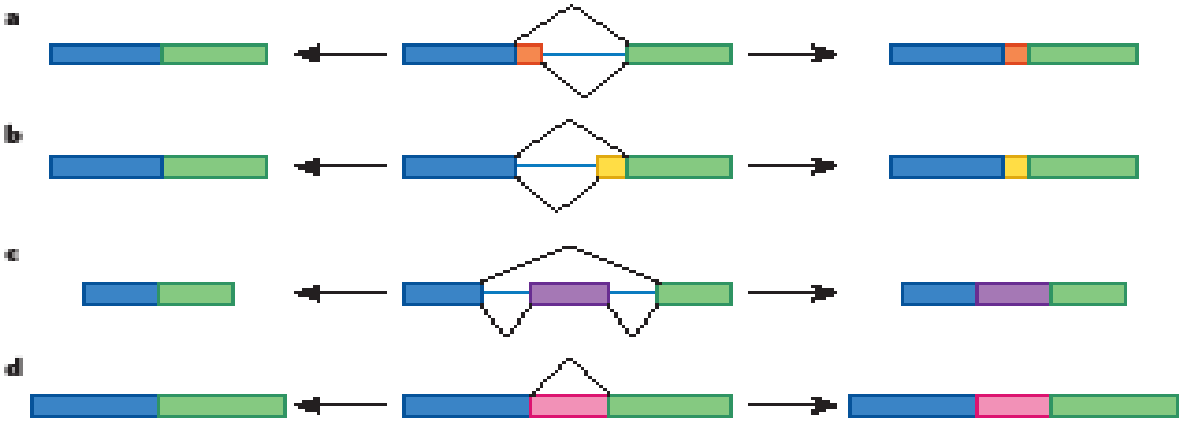


Figure 1-3. Four basic types of alternative splicing: alternative 5′ splice-site selection (a), alternative 3′ splice-site selection (b), cassette-exon inclusion or skipping (c) and intron retention (d). The rectangles in the center represent pre-mRNAs. For each pre-mRNA, the black lines span the regions that can be spliced out, with the lines above corresponding to the mature mRNA shown on the left and the lines below to the mRNA on the right.

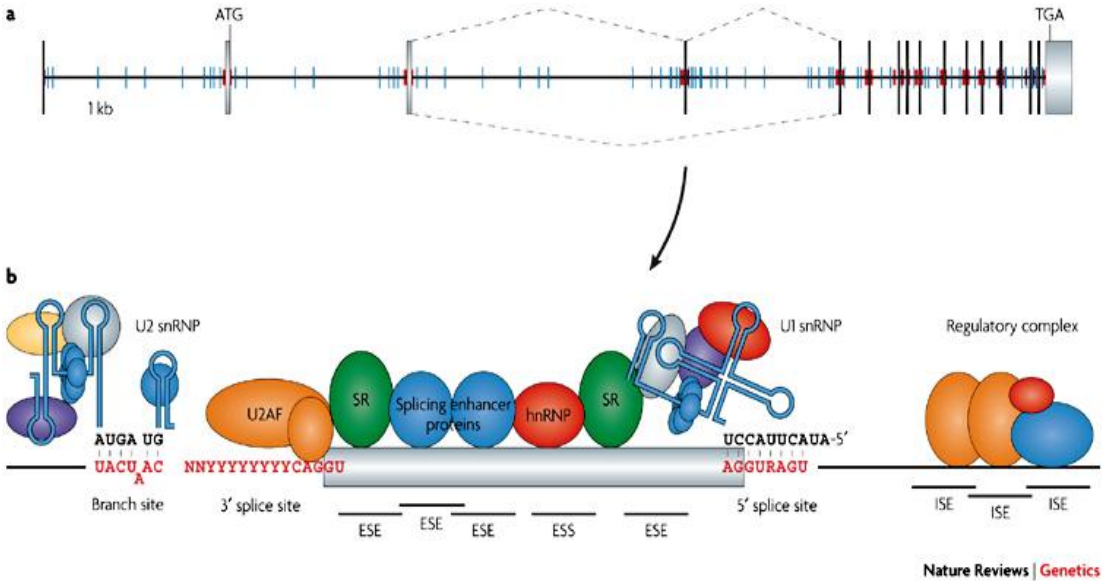


Figure 1-4. Splicing mediated by intronic cis-acting elements and trans-acting regulatory factors.

CHAPTER 2
BACKGROUND: STATE OF THE ART

In this chapter, I give an overview of methods used in alternative splicing studies, followed by an introduction to RNA-Seq technology. I then provide an introduction to existing computational tools to analyze alternative splicing using RNA-Seq technology. Finally, I list limitations of existing alternative splicing analysis tools.

Methods to Analyze Alternative Splicing

Traditional Experimental Approaches

Methods to analyze alternative splicing have improved in power in recent years, from experiments using low throughput experiments such as ESTs profiling, to exon-junction arrays [31].

DNA microarray technologies have been used extensively for high-throughput analysis of thousands of genes, in many applications such as gene expression analysis and association studies [46, 47]. Exon junction microarrays are widely used to study splicing and alternative splicing. Exon junction arrays use DNA probes made from sequences around splice sites on known exons to infer the presence of specific splice junctions. Overall, microarray technology provides simultaneous analysis of thousands of genes in a single experiment. However, microarray technology also suffers from several disadvantages. It depends on prior knowledge of the genes to be investigated. In addition, probe design problem such as the lack of binding specificity and probe cross-hybridization make it difficult to precisely interpret the hybridization signal [48, 49]. Furthermore, microarrays only detect the probe intensity signal (indirect measurement) of the gene expression, thus they are providing a less efficient measurement of the real gene expression [50].

RNA-Sequencing (RNA-Seq) Technology

RNA-Seq technology is increasingly popular as a tool to study the mRNA population in the cell. In RNA-Seq technology, mRNA is reverse transcribed to cDNA libraries which are amplified by PCR to produce many DNA short fragments as described in Figure 2-1 [51]. However, this reverse transcription may lead to amplification biases depending on factors such as RNA secondary structures. Subsequently, the amplified short DNA fragments are sequenced using high-throughput sequencing platforms. High-throughput RNA-Seq provides a large amount of information about the transcriptional state of the cell, under the form of short reads that map to genomic regions corresponding to gene expression [52]. The popularity of RNA-Seq is also due to the fact that the cost of deep sequencing has dropped significantly in the past few years.

Short DNA reads from RNA-Seq datasets mapping to the exons of a gene indicate that the corresponding gene is likely expressed. Exons and junctions that are specific to a particular isoform of a gene make it possible to identify the splicing isoforms that are expressed [53], and more accurate isoform discovery can be achieved by analyzing reads that map to exon-exon junctions. Finally, the number of reads aligning to a gene region, when normalized by the length of the region, provides an estimate of the expression level of that gene [54]. The results of an RNA-Seq experiment can therefore provide a snapshot of the RNA landscape in a population of cells in a given state, ranging from a catalog of all RNA molecules represented in it, to their exact structure, to their relative expression levels [46].

High-throughput RNA sequencing (RNA-Seq) is playing an increasingly important role in genome-wide alternative splicing analysis since it enables scientists to detect

constitutive and/or alternative splice junctions as shown in Figure 2-2 [54], and putative novel transcripts. Furthermore, it also helps us the study of expression regulation of transcribed non-coding RNAs, fused genes and functional microRNAs. Overall, RNA-Seq is playing and will continue to play an important role in studying post-transcriptional regulation because it does not require prior knowledge, and it offers single-base resolution and flexible analysis for different gene expression levels [54].

Existing Computational Tools for Alternative Splicing Studies

Bioinformatics is assuming an increasingly crucial role in analyzing the large datasets produced by next-generation sequencing technologies such as RNA-Seq. The rapid increase in the volume of available sequencing data creates a stronger need for robust computational analysis. Therefore, the implementation of accurate and efficient pipelines for RNA-Seq data analysis on alternative splicing is currently an important task for bioinformatician. Most computational tools to analyze splicing and alternative splicing using RNA-Seq data address one or more of the following three tasks: splice junction identification, *de novo* gene discovery and isoform expression estimation. All three tasks ultimately rely on DNA short read alignment with reference genome, as the first step to recover the presence of transcripts.

DNA Short Read Alignment using Bowtie

Bowtie is a short DNA sequence aligner that maps short read DNA sequences generated by many commercial platforms such as Illumina to a large genome in a fast and memory-efficient way[43]. Bowtie uses the Burrows-Wheeler transform (BWT) and the FM index to index the reference genome and it is able to handle short DNA sequence alignments against human reference genome in a typical workstation.

Splice Junction Identification

Junction identification plays an important role in characterizing alternative splicing regulation. Accurate prediction of splice junctions is a fundamental step for the correct assembly of gene isoforms and the subsequent expression analysis. RNA-Seq provides a convenient tool for junction identification because DNA short reads are generated from cDNA reversed transcribed from mRNA. Therefore, these spliced transcripts contain sequence from different exon sequences that allow us to infer the splicing junctions. There are a few splice junction prediction software packages available to analyze RNA-Seq data, of which Tophat is currently the most popular one.

Tophat

Tophat is an splice junction prediction package as shown in Figure 2-3 [44]. Tophat uses Bowtie to align short reads to the reference genome in the first phase. It then assembles splice junction through the ‘seed-and-extend’ method to assign unaligned reads to putative splice sites defined by canonical splice signals. Tophat is able to capture splice junctions previous reported in annotation, and it also allows detections of novel splice junctions.

De Novo Isoform Reconstruction

With the wide-spread application of RNA-Seq technology, there is a strong interest in *de novo* transcript reconstruction pipelines to analyze the global RNA content, especially for organisms lacking comprehensive gene annotations. The *de novo* discovery of transcripts using RNA-Seq data will give us a clearer picture of the RNA content with less bias compared with software relying on existing gene annotations.

Cufflinks

Cufflinks is a large-scale isoform assembly and quantification tool [55]. Cufflinks assembles the isoform structures using Dilworth's Theorem which is widely used to partition finite partially order sets into a minimum set of paths[56] . Subsequently, it estimates transcript abundances through a generative statistical model that aims to maximize a likelihood function built from the short DNA alignments in the isoform paths estimated from Dilworth's Theorem. Cufflinks is able to find previously unknown genes, especially for less studied organism, using RNA-Seq datasets with a relative large read length (such as 60bp or more).

Gene Expression Analysis

The estimation of isoform expression with RNA-Seq data is important in investigating mRNA population. Disruptions of alternative splicing may lead to the presence of abnormal isoform variants and/or altered abundance of gene isoforms that are associated with diseases. Altered ratios of isoform expression levels are a direct effect of some pathogenic changes [15].

MISO

MISO is a software package to estimate isoform expression levels using RNA-Seq datasets [40]. It incorporates the Bayes' rule and a likelihood model to characterize the assignment of reads to different isoforms. Therefore, MISO's estimation of isoform expression level is represented as a likelihood factor indicating the probability of finding a gene with the observed differential expression of its isoforms.

Limitations of Existing Tools

Existing splice junction detection pipelines, such as Tophat, suffer from drawbacks such as the requirement to use canonical splice signals. Furthermore, these

pipelines depend almost exclusively on sequence alignments and ignore biological features such as distributions of intron sizes and presence of Branch Point Sequences (BPS) [57], in finding splice junctions using RNA-Seq datasets. In addition, there are very few software packages available that perform de novo gene discovery from poorly annotated organisms [58]. Therefore, there is a need to improve existing methods to profile alternative splicing using RNA-Seq datasets.

Cufflinks is not capable of accurately reconstructing isoforms, especially for isoform with a large number of exons and when the RNA-Seq throughput is low as I will show in Chapter 5. The limitation of Cufflinks is that it only relies on short DNA reads mapping to reference genome to find new genes. However, a low RNA-Seq throughput will generate insufficient reads to accurately annotate exons and junctions. On the other hand, the proper inclusion of splice junction predictions is not used in the Cufflinks gene discovery algorithm. Therefore, more robust *de novo* gene discovery pipelines should be developed to analyze information contained in RNA-Seq datasets, especially on splice junctions and gene expressions.

MISO is unable to analyze multiple control and experimental samples simultaneously to find genes with differentially expressed isoforms. It relies on a single comparison between control and experimental datasets, so additional statistical analysis is required if the investigator is interested in studying differential expression using multiple control and experimental RNA-Seq datasets. Therefore, an expression estimation tool should be developed to handle biological replicates to reduce the post-processing work.

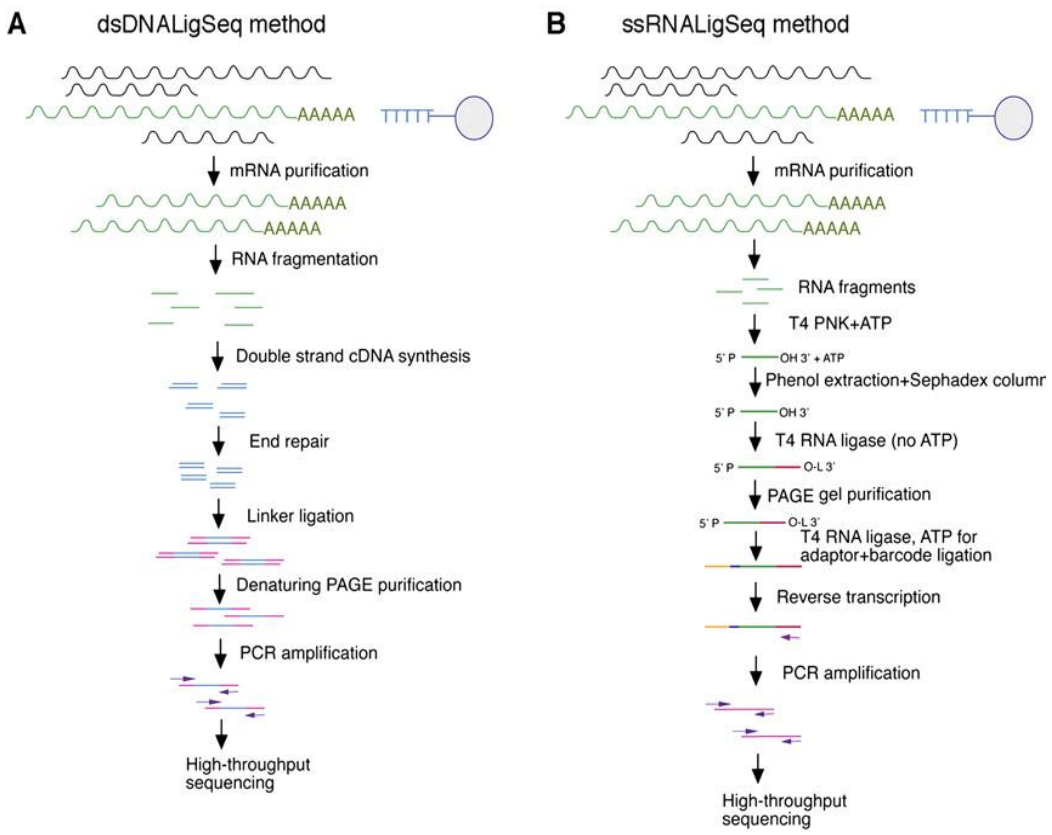


Figure 2-1. A brief description of the RNA-Seq technology.

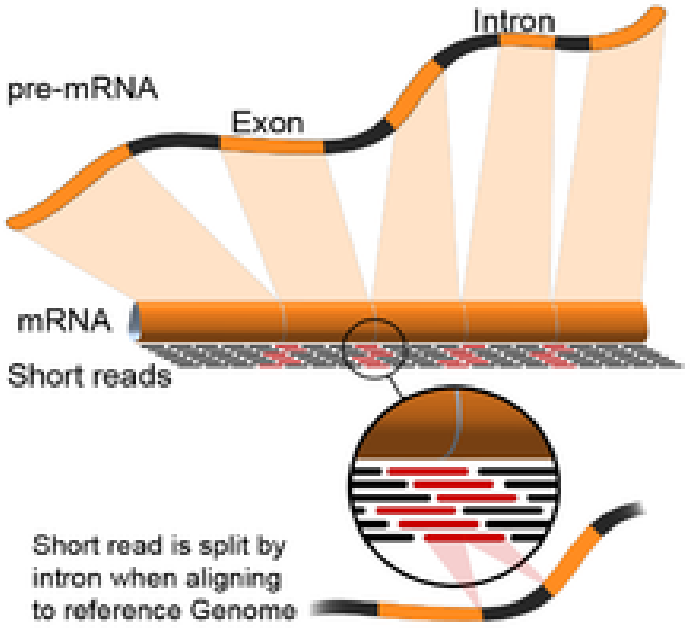


Figure 2-2. General principles of mapping short reads from RNA-Seq short reads to genomes.

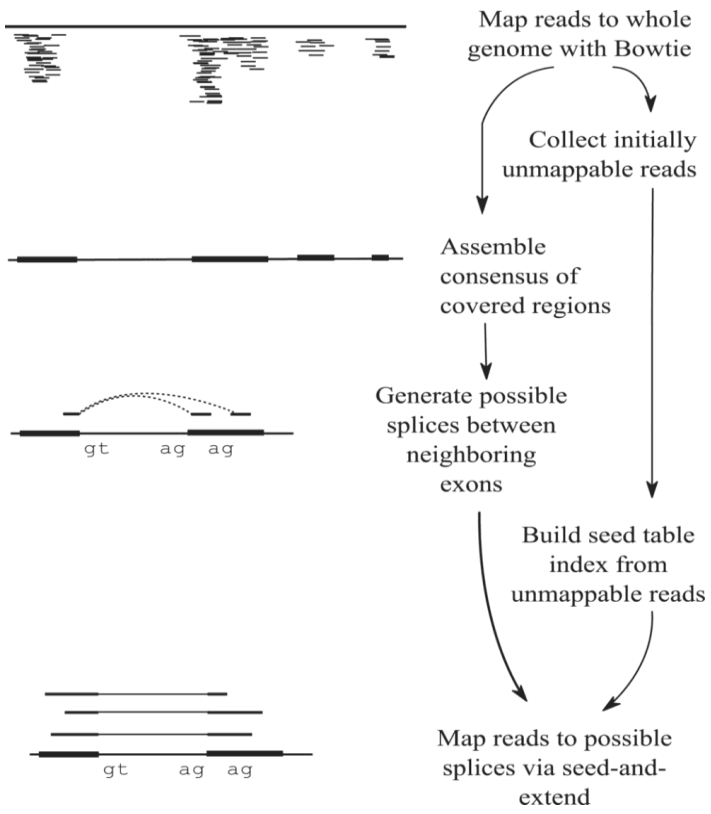


Figure 2-3. The flowchart of splice junction assembly in Tophat

CHAPTER 3
PROPOSED AIMS AND RESEARCH PLANS

This chapter begins with a discussion of the specific aims of this dissertation. I then explore approaches to analyze alternative splicing regulation and propose new methods to improve the understanding of alternative splicing using RNA-Seq datasets.

Specific Aims of This Research

Genome wide analysis of alternative splicing using computational methods has the potential to improve our understanding of diseases regulated by splicing factors. For example, genes from the Mbnl family are important splicing regulating factors that may cause the disruption of alternative splicing regulation in many other genes. The main purpose of this dissertation is to develop tools to study alternative splicing changes at the genome-wide level, including identifying splice junctions, obtaining alternative isoforms and identifying genes with differentially expressed isoforms, to improve our understanding of the splicing regulation.

Identify Alternative Splicing Junctions using Mbnl3 Isoform Knockout Mouse

In the first place, to study changes in splicing regulation in the *Mnbl3* isoform knockout mouse, I obtained a comprehensive annotation of splice junctions. The identification of splice junctions is a prerequisite in constructing splicing isoforms because splice junctions carry direct evidence of each isoform's structure. In this work, I am interested in identifying splicing junctions to detect alternative splice sites and cassette exons. Genome-wide profiling of splicing junctions can uncover a large amount of preliminary information regarding the transcriptome composition, which facilitates downstream analysis of transcriptional regulation and translational control. For example, alternative usage of splice sites or cassette exons may lead to a read frame shift, which

may produce altered protein sequences and as a consequence disrupt cellular functions.

Studying the Composition of the Transcriptome

Transcriptome composition is dynamic, and is affected by many factors, including tissue types, environmental changes, pathological conditions and developmental stages [15, 20]. *De novo* gene discovery reduces the bias toward the understanding of the transcriptome composition compared with analysis using existing gene annotations. It is also a robust approach to study differential transcriptional modifications and gene regulation in different developmental stages and/or tissue types. RNA-Seq technology offers an effective way to estimate transcriptome composition because short DNA reads are generated from the whole mRNA population.

Computing Gene Expression Levels to Study Expression Regulation

More than 80% of the genes from higher organisms such as mouse and human express multiple isoforms [31]. Studies have also demonstrated that the disruption of splicing may lead to the balance between different isoforms of the same gene to be modified, leading to changes in the production of proteins. In addition, even for genes expressing a single isoform, the change of expression level could lead to an increase or decrease of protein production. In this thesis, I developed computational methods to estimate absolute and relative isoform expression levels, which will lead to identifying genes with differentially expressed isoforms. The genome wide analysis of gene expression using RNA-Seq dataset can uncover important gene candidates with differentially expressed isoforms, which will increase our understanding of the underlying mechanism of splicing regulation.

Alternative Splicing Regulation Studies by Computational Analysis of RNA-Seq Datasets

Currently, there are few software packages designed for comprehensive genome-wide alternative splicing analysis [59]. Software pipelines such as Tophat, Cufflinks and MISO are designed for a specific topic I addressed previously. In this study, I am developing a new computational pipeline called PASTA for the analysis of alternative splicing through RNA-Seq.

Increase of Sensitivity in Splice Junction Detection

I am developing an analysis pipeline that includes an innovative algorithm for splice junction detection based on the following two facts. First, many pre-mRNAs in eukaryotic organisms undergo splicing to produce a mature mRNA containing multiple exons. Second, the majority of eukaryotic exons are typically longer than the short DNA read size produced by most experiments from RNA-Seq platforms. This ensures that short reads containing splice junctions are formed by joining sequences from separate exons, with one of them carrying a fragment of sequence from the end of one exon and another fragment of the short read sequence coming from the beginning of a second exon on the same chromosome region. After choosing a particular position in a DNA short read to split it into two sequences, the program is able to identify the two fragments each carrying sequences from one of the exons spliced together during the splicing process. As a result, identification of splice junctions is achieved without reliance on existing gene annotations.

Here I describe the first module of the pipeline, whose purpose is to identify all exon-exon junctions that can be inferred from aligned short reads, with high sensitivity and specificity. This is accomplished through a special patterned fragment identification

algorithm and a heuristic method to score junctions on the basis of biologically relevant features. This pipeline is designed to efficiently detect alternative splicing events, using a scoring scheme through logistic regression that measures the probability each splice junction is true according to existing gene information. This scoring system can effectively compute splice junctions given available short read information.

De Novo Gene Discovery to Characterize mRNA Population

It is estimated that 90% of genes in eukaryotic organisms contain multiple exons [1]. More than 98% of genes in human undergo alternative splicing, and the number of isoforms generated from each gene typically ranges from 2 to 20 [28]. Therefore, gene discovery pipelines should reflect the diversity of gene isoforms and the fact that most isoforms are composed of many exons.

Many *de novo* gene finding pipelines were developed over the past two decades. A few pipelines rely on canonical splice signals that are used by the majority of genes consisting of multiple exons [60] Alternative splicing is emerging as a universal process that increases the transcriptional complexity and diversity from a relatively small number of genes. Therefore, alternative splicing generates a large number of transcripts with distinct features such as the usage of non-canonical splice sites and the presence of trans-splicing [60]. However, many existing gene annotations are still insufficient to effectively represent comprehensive gene information, which are highly dynamic depending on different developmental stages or tissue types [59].

In this work, I propose the second PASTA module for *de novo* isoform reconstruction, able to effectively discover genes from DNA short reads that are mapped to reference genome. In addition, PASTA actively utilizes codon utilization bias since different organisms have different preferences in utilizing codons for making

amino acid sequences [61]. This bias is also manifested by the preferences in the usages of synonymous codons [62]. Therefore, a proper inclusion of codon utilization information could be an effective way to improve the prediction accuracy when discovering coding transcripts.

Estimation of Expression Levels of Genes with Differentially Expressed Isoforms

Many genes in higher organisms express multiple isoforms. The disruption of the relative expression levels of different isoforms is associated with diseases such as Spinal Muscular Atrophy (SMA) [63]. In this work, I estimate isoform expression levels using a Genetic algorithm, which is based on an optimization process of alternative exons that are the direct manifestation of particular isoform expression. The description of this algorithm is provided in Chapter 4.

Most commercial platforms enable the sequencing of multiple samples in a single RNA-Seq experiment. As a result, biological studies frequently use multiple control and experimental groups to monitor the altered expression changes in a disease sample compared with a control. Therefore, a robust transcriptome profiling pipeline should fully capture the subtle expression changes within multiple samples. Unfortunately, most tools such as MISO perform differential gene expression estimation using pairwise comparison from two samples in each analysis [64]. In contrast, new tools need to be developed to examine multiple control and experimental samples simultaneously of isoform expression. In addition, the visualization of gene expression estimates by a figure containing the expression levels from individual "gene signatures" such as exons or splice junctions may provide an elegant way to look for differentially expressed isoforms for a small number of genes of interest. These new features, which will play an

important role for biological community, are absent in many RNA-Seq data analysis tools such as MISO.

Studies on Relation of Alternative Splicing and Disease using Mbnl3/Mbnl2

Previous studies have shown that mutations in the *Mbnl1* gene are strongly associated with the onset of DM. However, two other related genes *Mbnl2* and *Mbnl3* are relatively less studied, and evidence of their role in disease onset is unclear [65]. In DM disease, expression levels from members of the Mbnl gene family are down regulated which may lead to a subset of pre-mRNAs to be aberrantly spliced [66].

Mbnl2 is an important gene that is dysregulated in DM and *Mbnl2* knockout mice show disease phenotypes associated with DM syndromes [67]. In this study, I have used the double knockout mice with exon 2, which encodes the initiation codon form *Mbnl2* protein, removed by a homologous recombination strategy.

Mbnl3 is another important member of the Mbnl gene family that inhibits muscle differentiation and may function directly or indirectly to DM-associated muscle degeneration [68]. Two common isoforms are commonly observed in Mbnl3. One of the isoforms encodes a 37 kd protein, and the other isoform encodes a 27 kd protein (with exon 2 excluded from this isoform). In this study, in-vitro cell lines with *Mbnl3* isoform knockout (exclusion of constitutive exon 2) are sequenced in order to study global splicing regulation. The purpose of our study is to investigate the genome-wide splicing changes as a result of the *Mbnl3* isoform knockout. I am going to conduct a systematic analysis of alternative usages of splice sites, cassette exons and alternative polyadenylation to draw a complete picture of alternative splicing profiling using transcriptome sequencing samples.

A Better Understanding of Alternative Splicing in Maize

PASTA is not limited to alternative splicing analysis using RNA-Seq data from mammals, but can process RNA-Seq data from any organism in which splicing is observed. Because of the economic importance and its high quality reference genome, maize is used widely in alternative splicing analysis in studying the regulatory mechanisms of improvements in yields. In this work, I use PASTA with maize RNA-Seq datasets to conduct genome-wide alternative splicing analysis, including the identification of alternative isoforms and characterization of putative transcripts not appearing in annotations. The results could be important to improve our understanding of splicing regulations in plants such as maize.

PASTA Features and the Comparison with Existing Tools

Table 3-1 shows a comparison of PASTA features against similar existing software tools. I will provide detail methods and algorithm for these features in Chapter 4.

Table 3-1. Alternative splicing features PASTA provides in comparison with other tools.

	PASTA	Tophat	Cufflinks	MISO
Minor splice sites	Identify novel splice sites.	Feature not included in the first release.	NA	NA
Trans-splicing	Analyze trans-splicing by default.	NA	NA	NA
Isoform reconstruction	Use codon utilization to optimize isoform reconstruction.	Construct exon clusters.	Use graph theory to minimize number of isoforms reported.	NA
Differential expression	Offer isoform-approach using Genetic algorithm and exon-approach using expression fold changes.	NA	Estimate transcript abundance by short read counts.	Use Bayes model to assign reads to each isoform. Perform pairwise comparison between two samples.
Graphical display	User-friendly figures to visualize differential expression.	NA	NA	NA

CHAPTER 4
METHODS AND PROCEDURES

In Chapter 3, I introduced the research plan of this work. In this chapter, I provide the implementation details of a software package for alternative splicing analysis using RNA-Seq datasets. The software package is implemented in three modules, each of which is targeted on a specific analysis type, as described in Chapter 3. The detailed algorithms and methodologies of the three modules will be addressed in order.

RNA-Seq Dataset Simulations using RNA-Seq Simulator

In developing the algorithm and evaluating the performance of PASTA in splice junction detection, I have implemented a RNA-Seq simulator that generates RNA-Seq short reads using any given reference genome and gene annotations.

In this study, I have generated single end DNA short-read sequences (in FASTQ format) with a read length of 50 nucleotides. Quality scores in Illumina format (GA Pipeline version <1.3) are randomly simulated for each read with a normal distribution. The detailed simulation procedure is described in the following. In the first place, I randomly choose a fraction of existing annotated genes from the organism of interest (I used mouse gene annotations in this study). For genes with alternative isoforms, I choose a random subset of the isoforms for use. Second, I generate the transcript sequence from the chosen isoform by removing intron sequences and concatenating the exon sequences. Finally, a specific number of short read sequences of 50 nucleotides in length are generated from random positions in the chosen isoform, such that the isoform expression level is approximately equal to the desired expression value, specified in the simulator in terms of Reads Per Kilobase per Million reads (RPKM). To make the data more biologically meaningful, I also introduced random sequencing

errors and Single Nucleotide Polymorphisms (SNP) at a frequency of 1/100 and 1/500 respectively.

The RNA-Seq data simulator enables us to generate datasets that are essential for calibrating the software performance. First of all, each short-read sequence is tagged with a label which tells the origin of the read, including the isoform it is produced from and its exact location on the chromosome. Second, I can manipulate gene expression levels and simulate differential expression of isoforms, which gives us flexibility to study different experimental situations and improve the subsequent software implementation.

Splice Junction Prediction using Patterned Alignments of RNA-Seq Data

The first step in the PASTA is to align the RNA-Seq short reads to the reference genome. This step is usually performed using an existing, fast alignment tool like Bowtie [43] and MAQ [69]. Since DNA short reads are aligned to the genomic sequence, reads that are entirely contained within exons will align correctly, while reads falling over the junctions between two exons will, in general, fail to align. The main task performed by PASTA is to infer the exact location of exon-intron boundaries using the unaligned reads.

In contrast to the "seed and extend" method used by the majority of similar programs [55], PASTA relies on patterned alignments combined with a logistic regression model. The purpose of the logistic regression model used in PASTA is to score each putative intron-exon junction according to its biological “context”: for example, the presence of canonical splice signals [11], the expected distribution of intron sizes, the presence of regulatory elements such as the Branch-Point Sequence (BPS).

A Logistic Regression Model for Splice Junction Prediction

Many model organisms have intron sizes that span several orders of magnitude, from less than 100 base pairs to over 100,000. However, the intron size distribution follows a characteristic curve, since shorter introns are more common and longer introns are rare. This can be modeled by a Pareto distribution whose parameters are organism-specific.

The junction prediction algorithm assigns a weight to each predicted junction, using a logistic regression model that takes into account the factors that characterize its context, such as the presence or absence of canonical splicing signals, the posterior probability of the intron size from the Pareto distribution, and the alignment quality, as described in Figure 4-1. The logistic regression model can be written as $f(\mathbf{Z}) = \boldsymbol{ß}_0 + \boldsymbol{ß}\mathbf{Z} + \varepsilon$, where $f(\mathbf{Z})$ is the logistic regression value, $\boldsymbol{ß}_0$ and $\boldsymbol{ß}$ are coefficient vectors and $\mathbf{Z}$ is a vector containing the values of regression factors described above. The coefficients of the logistic regression model and of the Pareto distribution are estimated from existing splice junction annotations for the species under consideration [70]. I provide pre-computed models for many commonly studied organisms.

The logistic regression model is used to score the “pattern-pair” alignment matches and identify the best combination of alignment positions for the two fragments in each pair. The score is computed as $1/\left(1+e^{-f(z)}\right)$ resulting in a value in the range 0 to 1, with negative values of $f(z)$ producing probabilities close to 0, and large positive values producing probabilities close to 1. All aligned pairs whose score is above a threshold set by the user are retained and clustered based on their position, producing a predicted junction. PASTA is also able to deal with paired-end reads: the two reads in

each pair are initially aligned separately as described above, and information about the position of the two reads is incorporated into the logistic regression model to increase the accuracy of the splice site predictions.

The branch point sequence

One of the critical steps in RNA splicing is the formation of the lariat structure, which is generally anchored to conserved sequences 10 to 50 nucleotides upstream of the 3' splice site. The BPS is highly conserved in yeast with the consensus sequence UACUAAC , In human, the BPS is less conserved, but studies reveal that human BPS can be represented as yUnAy where 'y' stands for nucleobase C or T [71].

The polypyrimidine tract

The PPT plays an important role by acting as the binding recognition site for pre-mRNA splicing factor *U2AF(65)* during the early stage of splice site recognition. However, the PPT is less conserved in higher organisms compared with lower organisms such as yeast [72].

Logistic Regression Model Training Using Existing Annotations

To use the logistic regression model to generate the likelihood score in splice junction prediction, I implemented a method to use existing annotated genes to obtain the logistic regression coefficients for the organisms under study. Information including the individual intron size, presence of canonical signal and branch point sequence from given annotated genes is recorded and modeled by statistical software R using the logistic regression procedure Generalized Linear Model (GLM). Finally, a probability score is computed for a new junction prediction using the coefficients obtained in the model training step.

Pareto distribution approximation with introns from gene annotations

In most of the organisms that contain spliced genes, the sizes of introns typically follow a certain trend that is different from other organisms. However, there is a general rule in which shorter splicing introns are preferred over longer introns. This behavior can be captured by the Pareto distribution which is first proposed to model the distribution of income that a small number of people own a large portion of the wealth.

Pareto distribution to model intron distribution from annotated genes

In this study, intron sizes from NCBI RefSeq genes are numbered by ascending order. I record the occurrences of intron in each region size. The recorded occurrences for each region will be further transformed by logarithm so the intron occurrences will decreases steadily as their size in logarithm increases. Finally, I rank the intron size by the formula $r_i = 1 - x_i / t$, where x_i represents the logarithm of the number of intron occurrences in the ith intron and t is the total number of introns under study.

The NCBI reference sequence

The NCBI RefSeq database contains a comprehensive set of annotations describing experimentally validated genes and providing exact genomic coordinates for transcripts and exons. It defines a uniform set of standards to present information such as gene characterization and genome annotation [73]. Complete genome data from a wide range of species, including the widely studied model organisms are available from RefSeq.

Patterned Alignments

Generation of patterned split fragments

The PASTA algorithm considers each unaligned read, in turn, and generates two sets of “patterned” subsequences from it, by splitting it at different cutoff points.

Denoting the read length with *n* and a stepping distance *s* (4 by default) and a minimum fragment size *m*, each patterned pair will consist of the sequence from the start of the read to position *p* and of the sequence from position *p* to the end of the read, where *p* ranges from *m* to *n-m* in steps of *s* (Figure 4-2). For example, if *n*=36 and *m*=6, each read will produce seven patterned pairs of lengths (6, 30), (10, 26), (14, 22), (18, 18), (22, 14), (26, 10), (30, 6). All these fragments then undergo a second round of alignment. Fragments longer than a set minimum size (16 by default) are again aligned to the reference sequence with Bowtie, while each remaining short fragment is aligned to the region around the other fragment in the pair, using a string matching algorithm or dynamic programming sequence alignments.

In this example, I set the minimum alignment size to 14 and step size to 4. Patterned pairs (14, 22), (18, 18) and (22, 14), as well as single fragments (-, 30), (-, 26) and (26,-), (30,-) (where '-' indicates either the left or right fragment not in the whole genome alignment) will be aligned to reference genome. To identify exon-intron boundaries, PASTA looks for alignment matches from all patterned pairs (14, 22), (18, 18) and (22, 14) in the first place. If alignment matches are reported for both left and right fragments in a given patterned pair, PASTA will optimize the exon-intron boundary adjusting the left and right fragments around the cutoff point. Suppose alignment matches are found for patterned pair (14, 22). I will slide the cutoff point by 1 or 2 nucleotides in both directions, resulting in alternative patterned pairs of (12, 24), (13, 23) and (15, 21), (16, 20). The optimized exon-intron boundary is the patterned pair that minimizes the sequence alignment mismatches. In case of ties, an arbitrary patterned pair will be chosen. If alignment matches are found for multiple patterned pairs for the

same read, PASTA fits a logistic regression model for each reported patterned pair and picks the final patterned pair with the highest probability score.

Local alignment to find short fragment matches

There are cases in which an alignment match is only observed for either the left or right fragment in a patterned pair. In this situation PASTA picks the fragment with the longest aligned match, and performs a local alignment step in order to determine the location of the other half of the pair (Figure 4-3). Suppose for example that the left fragment (26,-) is the longest fragment observed to have an alignment match. PASTA will search the chromosome region adjacent to its position to locate the optimal position of the remaining right fragment of length 10. The local alignment can be done in two ways. If the user prefers the alignment to be fast, PASTA will use exact string matching. In the other case, PASTA will use the Smith-Waterman algorithm that employs a dynamic programming approach to compute the optimal local alignment allowing insertion or deletions.

Smith-waterman dynamic programming algorithm

Smith-Waterman is a dynamic programming algorithm that is guaranteed to find a local alignment using a specified scoring system. For short DNA fragment alignments, the scoring scheme typical includes the scores for alignment match, mismatch and gaps [74]. This algorithm is based on the dynamic programming property that the optimal alignment is composed of an optimal alignment with a sequence of 1 nucleotide shorter, and the decision with regarding the last nucleotide that maximizes the overall score.

Optimization of pattern alignments

The size of the region analyzed by the local alignment procedure can be configured by the user, with a default of 100,000 base pairs. This step obviously

increases the computational cost of the algorithm, since each short read gives rise to a large number of fragment pairs; on the other hand all these fragments are relatively short and can therefore be aligned very efficiently. On average, the second round of alignment takes 2 to 5 times longer than the initial one, but this performance penalty is compensated by a large gain in accuracy, as shown below. Also in this case, if multiple alignment matches are observed, PASTA will select the alignment that produces the highest probability score after applying the logistic regression model to the intron it predicts.

Junction identification

Patterned alignments enable PASTA to find a putative splice boundary for each patterned pair. PASTA will combine these splice boundaries to construct the final splice junctions (Figure 4-4). First, if a single splicing boundary is reported in any specific chromosome location, PASTA directly reports it as the final splice junction. Otherwise, if multiple splice boundaries within a range up to 3 are found, PASTA will report the median of these splicing boundaries as the final splice junction.

Junction accuracy by false positive rate and false negative rate

In order to evaluate the junction predictions, I have compared PASTA predictions with existing gene annotation. The accuracy of PASTA predictions was evaluated in terms of false positive rate and false negative rate. False positives FP are defined as predicted splice junctions that are not originally found in the simulated data. False negatives FN are splice junctions that are originally found in the simulated database but are not predicted by the algorithm. The true positives TP are the junctions that software package correctly predicted and are found in the simulated junctions. The true negatives TN are those reads that are not simulated junctions, and are not reported by

the software as well. In general, I have false positive rate $FPR = FP/(FP+TN)$ and false negative rate $FNR = FN/(FN+TP)$.

Paired-end RNA-Seq strategy

PASTA performs patterned alignments on one or both ends of the paired-end reads for which no full matches are found. However, paired-end reads are more informative in terms of relative alignment positions from the two ends. In paired-end RNA-Seq data analysis, PASTA records the mapping location of the longest patterned fragments for both ends. The mapping location marks the chromosome location around which the new predicted junctions must be located.

De Novo RNA Isoform Reconstruction

I describe a reference genome guided isoform reconstruction module that finds gene isoforms from short DNA sequences on RNA-Seq experiment without relying on existing gene annotations. This PASTA module provides an innovative tool to study the transcriptome landscape with little or no prior knowledge on existing gene annotations. It is able to assemble gene isoforms from RNA-Seq data or splice junctions produced from junction prediction program. Finally, it includes a likelihood model to optimize the prediction accuracy of coding transcripts if preliminary codon utilization information is available for organisms under study.

Assembly of Novel Transcripts

RNA-Seq data are aligned to reference genome by PASTA aligner to identify exonic read mapping and splice junctions. PASTA then uses splice junctions to reconstruct initial RNA isoforms. Gene characteristics including exon size, junction length and intergenic span size used for defining RNA isoforms are parameterized for

the organism under study and are configurable by user. Finally, PASTA uses genetic algorithm to assign estimated expression levels to each RNA isoform.

Initial isoform reconstruction

PASTA uses predicted splice junctions generated from the splice junction aligner or other software such as Tophat to construct the initial RNA isoforms' exons. It scans the reference genomes for species under study to identify transcribed RNA isoforms where mapped reads and/or splice junctions are reported.

In reconstructing gene isoforms, PASTA orders the splice junctions according to their ascending chromosomal positions from reference genome and constructs the exons of predicted RNA isoform path from the splice junctions that constitute a valid gene satisfying initial gene definition. The steps of reconstructing RNA isoforms are as follow. For a new splice junction that appears either at the beginning of a chromosome or immediately following an already assembled RNA isoform, PASTA looks for transcriptional start site for this putative new transcript's first exon. If mapping coverage is not sufficient, for example, a read depth (the average number of times each nucleotide is sequenced) of less than 1 is observed for the first exon, PASTA will assign a size of 50 nucleotides to the initial exon, which is much smaller than the typical exon size in metazoan organisms, to prevent inclusion of intergenic regions. Subsequently, PASTA looks for the next exon following observed splice junctions and this procedure continues until the conditions defined for a valid RNA isoform is violated. In the meantime, PASTA keeps track of splice junctions that could not be included in the current transcript and initiates the same procedure of transcript reconstruction from these splice junctions, reporting the assembled transcripts if no exons are overlapping with the existing transcript predictions.

Alternative isoforms

PASTA uses a greedy approach to generate initial transcripts given the alternative splice junctions that are observed. When two splice junctions share the same 5' or 3' splice site, PASTA chooses the splice junction that produces the smaller exon size, aiming to reduce the risk from improperly skipping exons falling in between. By doing this, PASTA is able to generate a minimum number of initial transcripts as distinct gene clusters as shown in Figure 4-5. The alternative splice junctions not included in the initial transcripts are not discarded. Instead, alternative RNA isoforms with be generated from initial transcript from these alternative spliced junctions after all initial transcripts reconstructions are complete.

Optimization by codon utilization preference

If multiples ORFs are produced from RNA isoforms, PASTA only keeps one final ORF that has the highest quality using the following two approaches. First, if prior knowledge of codon utilization frequency from existing annotations for the organisms under study is available, PASTAs select the transcript having the highest likelihood score from codon usage maximization (described in next section). In the second approach, PASTA selects the transcript with the longest translated peptide sequences.

Prediction of Novel RNA Isoforms using Codon Utilization Frequency

To study codon usage preferences, I compute the difference between the observed codon usage frequencies in the ORFs of the gene under study with existing codon usage frequency for the organism of interest (Figure 4-6) [75]. A likelihood score of the ORF is derived with respect to the reference codon usage frequency. Let $f(c)$ represents the recorded frequency (probability) of codon c observed in the genes of the species I am studying. Given a RNA transcript composed of consecutive

codons $S = c_1 c_2 ... c_m$, codon usage probability is calculated *by* $P(S) = f(c_1) f(c_2) ... f(c_m)$. Assuming a random model of coding DNA, a codon usage of 1/64 is reported for every codon. The logarithm likelihood of codon usage reported for a particular amino acid sequence S will be $L(S) = \log\left[P(S) * 64^m\right]$ where m is the number of consecutive codons [75–77]. PASTA generates ORFs for each isoform that is constructed from the initial transcript, it computes likelihood scores for each peptide sequence translated from the ORFs, and finally selects the candidate with best likelihood score.

Gene Discovery using Existing Annotations

RNA-Seq datasets can also be used for isoform reconstruction in widely studied and well annotated organisms, such as mouse. PASTA uses existing gene annotations to look for alternative isoforms or novel genes as described in Figure 4-7. In the annotation guided module, PASTA uses alignment information, including exon mapping from short reads and splice junctions predictions, to validate the transcription of existing genes. New spliced isoforms are constructed from alternative junctions of any annotated genes. Transcripts found in un-annotated regions are labeled as *de novo* predictions using the strategy described as in the previous section.

Functional Annotation of Putative Transcripts using BLAST

In order to look for protein homology matches to identify important function domains of newly discovered genes, PASTA uses the BLAST alignment procedure. Protein sequences are aligned to the NCBI “non-redundant” sequence database. Homologous proteins with high identity will be recorded for further analysis on protein functions. In general, homologous sequences may indicate a sharing of common functional domains. In other words, I could predict the putative cellular functions of a

newly identified protein using information from homologous proteins. However, homolog gene searches may be inadequate in certain cases, especially if the “new genes” have undergone substantial modifications that lose their original coding potential. Therefore, an alternative approach in addressing these unknown genes is through a comparison with known pseudogenes. PASTA will compare the predicted transcripts with pseudogenes in computing the difference between them to look for potential protein sequence changes.

Transcriptome Studies using RNA-Seq Data

Defining Gene Coverage from Alignments of RNA-Seq Datasets

The term ‘expression level’ of a gene may be ambiguous in the context of gene expression because of different definitions. In this work, "expression level" of any transcript/exon is defined as the average times each nucleotide of an exon appears in a read. This definition gives us the absolute value of the expression estimate. Normalization of expression level among samples can be performed by dividing the absolute value by the total number of reads available for the sample studied with respect to a standard number of 10 million reads (this number is in the same order of magnitude of current commercial RNA-Seq throughput).

Estimating Isoform Expression Levels

Many genes from eukaryotic organisms are expressed in multiple isoforms. The relative expression levels of the isoforms are strictly regulated in different developmental stages and tissue types. RNA-Seq technology enables us to estimate individual isoform expression levels from read mapping information. However, many isoforms share a large proportion of the exons. To adequately estimate individual isoform expression levels from digital counts, I need a model to effectively utilize the

read count information from the alternatively spliced exons or exon portions, along with junction predictions. In this study, I adopted an approach based on Genetic Algorithm that aims to estimate isoform expression levels according to the process described in the next paragraph.

The Genetic Algorithm to Estimate Isoform Expression

In this model, an equal expression value is initially assigned to each individual isoform. The expression level of each exon will be the sum from the expression values of the isoforms that contain it. In the case of "private exons", for example exons that appear only in a certain isoform, the exon expression level recorded from mapping information will reflect the actual isoform expression level. The genetic algorithm applies iterative optimization steps by adjusting isoform expression levels to minimize the squared errors between the recorded exon expression levels and the predicted levels from the assigned optimization values (Figure 4-8).

In case of alternative splice sites, PASTA divides the alternative exon into two portions, with one portion shared by alternative isoforms and the other portion owned by a specific isoform. This exon division approach transforms the exon into two parts, a constitutive exon and a cassette exon, that can then be analyzed using the method described in the preceding paragraph.

Let $\vec{y} = [y_1, y_2 ... y_n]$ be the vector containing individual isoform expression values to be optimized (n is the total number of isoforms for the gene under study), and $f(i, j)$ be a binary variable representing the value of 1 in presence and 0 in absence of exon i in isoform j ($i = 1,2...m$, $j = 1,2...n$) where m is the total number of exons under

study, and support the average expression value for exon i is x_i. The optimization process will minimize the square error by:

$$\sum_{i=1}^{m}\left[x_i - \sum_{j=1}^{n}\left(f(i,j) * y_j \right) \right]^2 \qquad (4\text{-}1)$$

Differential Expression Analysis in Experimental/Control Studies

Most RNA-Seq platforms provide a way for multiple samples to be sequenced in a single run. As a result, biologists can design multiple biological replicates for control and experimental studies. Therefore, it is necessary to explore computer pipelines that enable simultaneous analysis of differential expression in multiple samples.

In this work, I have modeled the differential expression of gene isoforms on the basis of the expression levels of individual gene features such as exons and splice junctions, and in particular alternative exons and/or splice junctions, are computed. PASTA computes the expression level fold changes for each individual gene features, including exons and junctions. This approach is similar to microarray analysis where fold changes of probes are computed to infer estimated isoform expression changes. The result is a fold change value obtained by dividing the average expression levels in the experiment and the control.

PASTA computes the expression level for each exon of a gene under study by summing the total number of nucleotides from reads that are mapped to the exon region, and then dividing this number by the length of the exon. The expression level of a splice junction is the number of reads covering the splice junction observed in the prediction. However, there are cases in which the expression level of an exon is dramatically different across samples. Therefore, PASTA includes the computation of

exon expression standard deviation, to evaluate the fluctuation of exon expression across samples.

Parametric/Non-Parametric Statistical Tests

In finding genes with differentially expressed exons and/or junctions in multiple samples, I compute statistical significance using the Student t-test, and the Kolmogorov–Smirnov test (K–S test) to find exon and/or junctions with a statistically significant difference in mean values between the control and mutant samples. Student t-test is based on the normality assumption under the Central Limit theorem. In our analysis of *Mbnl3* datasets, a total of 5 control samples and 8 mutant samples were generated from *Mbnl3* datasets in Run1 and Run2 (described at appendix), so the sample sizes are relative small and may cause the statistical power to drop. Therefore, I also use the nonparametric statistical K-S test to test whether the control and mutant samples follow the same probability distribution function, and thus determine whether the control samples and mutant samples have an equal sample mean.

Defining Fold Changes between Mutant and Control Samples

The fold change is the average expression level of a gene feature in mutant groups divided by the corresponding value from control groups, and is undefined when expression level on either control samples or mutant samples is 0.

Machine Learning Algorithm to Select Alternative Isoforms of Interest

After features of gene models described in the previous sections are computed, PASTA stores the following information to a database to facilitate fast retrieval and re-computation. First, individual exon and splice junction expression levels for each biological replicate (or by RNA-Seq lane) are computed. Second, PASTA merges the exons from each gene's alternative isoforms into a non-redundant set of unique exons,

from which the original isoforms can be reconstructed. This procedure creates a set of standard unique exon coordinates that facilitates expression level analysis. Finally PASTA specially marks the exons and/or splice junctions whose fold changes is greater than a cutoff value specified by the user, and uses this to detect for potentially differentially expressed exon and/or junctions.

The filtering and optimization processes described above enable PASTA to group genes with differentially expressed isoforms according to different rules. First, PASTA computes two average gene expression levels for control and mutant samples respectively. A standard shift index SI, which describes the relative fold change of the mutant sample (or samples) with respect to control sample (or samples), is computed from average gene expression level of mutant sample (or samples) divided by control sample (or samples).

Next, PASTA computes the deviation index DI for each alternative spliced exon by dividing the average exon expression level of the mutant samples over control samples. PASTA then defines the changing index CI as $CI = DI / SI$. If a particular transcript is alternatively regulated in the experimental sample (or samples) versus the control sample (or samples), I will expect to see an opposite change effect of DI and SI. PASTA implements a scanning algorithm that will preferentially select genes with a large CI as candidate targets of differential regulation. Similarly, PASTA can also use splice junctions as supplementary evidence for identifying genes with differentially expressed isoforms. Splice junction expression level changes are computed similarly as in exons described above. In this study, we use both exon and junction expression changes to infer the differential expression of isoforms.

Finally, PASTA generates a graphical representation of isoform structures and expression signatures. Exon expression levels are plotted above the horizontal axis while exon junction recorded numbers are displayed below the horizontal axis. The expression levels on gene features for control and mutant sample (samples) are drawn in with blue and red color respectively to allow users to easily inspect the difference in overall gene expression and individual gene features. The purpose of this feature is to give the user an option to view the exon expression level and junctions in a distinct way.

Pseudo code of the genetic algorithm to compute isoform expression values

- Assign initial expression value of i (e.g i=10) to m (m>0) isoforms of the gene
- Loop for a certain number of times n (e.g n=100)
- Choose one of the m isoforms at random and change its expression level
- Change this isoform expression level either by 50% up or 50% down, then compute the new squared error of the isoform expression.
- If the new squared error is less than the existing score
- Save the change of expression to the chosen isoform, and assign this new score as existing score
- Otherwise, disregard the change

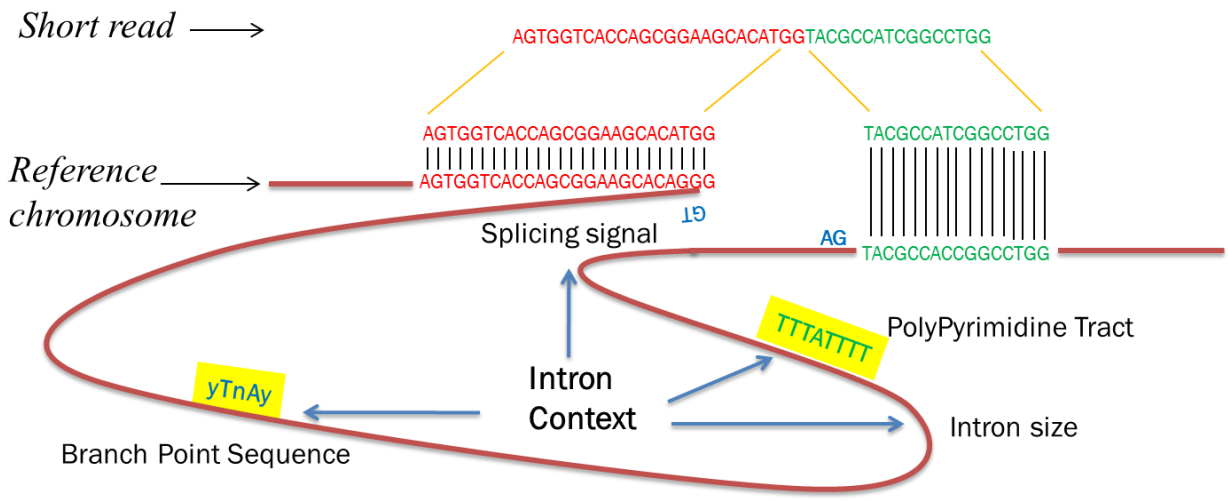


Figure 4-1. PASTA logistic regression model from intron context.

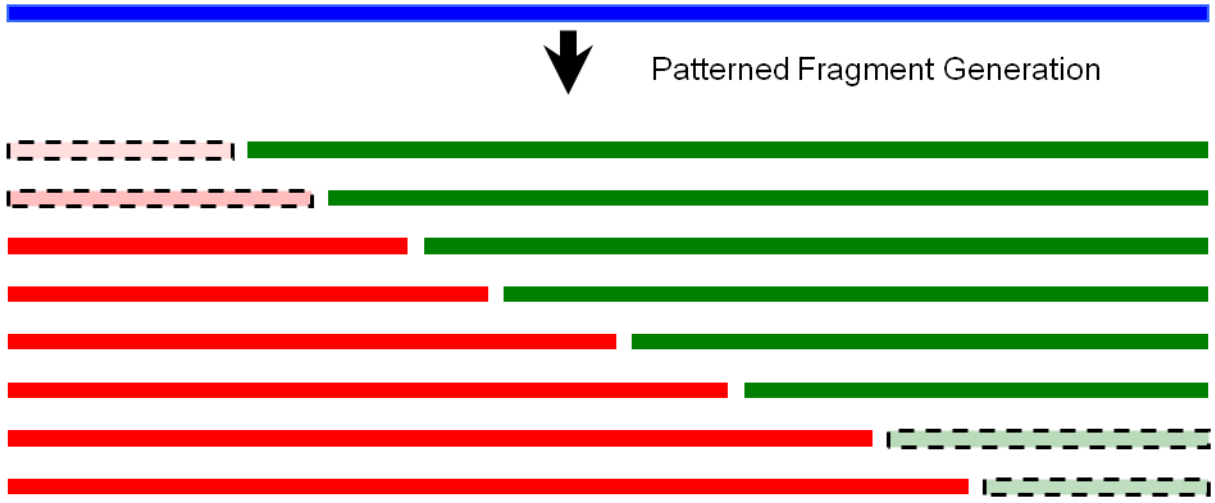


Figure 4-2. Flowchart on the generation of patterned sequence fragments from a RNA-Seq short read. The short fragments with dashed outline will be subject to local alignment defined by the chromosome region where the other longer fragment is uniquely aligned to.

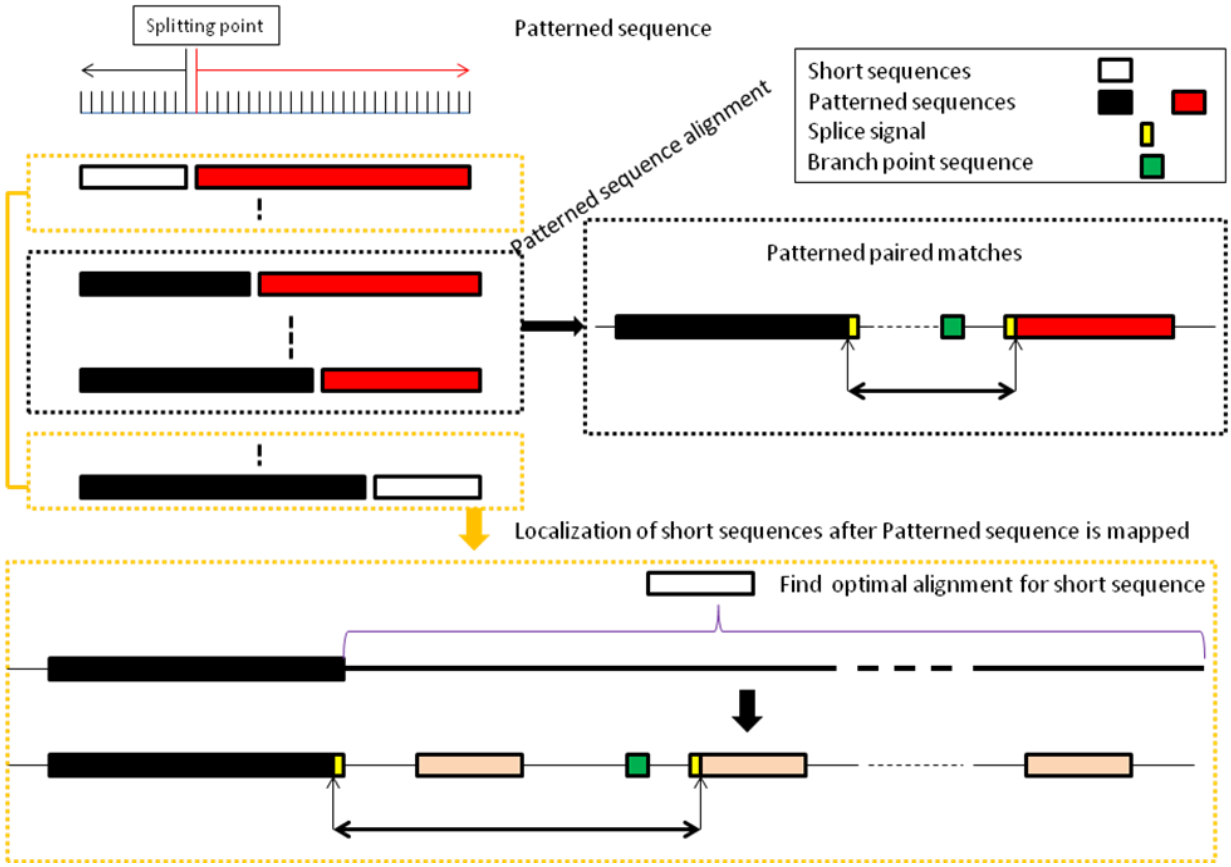


Figure 4-3. Flowchart of patterned fragment generation from short reads.

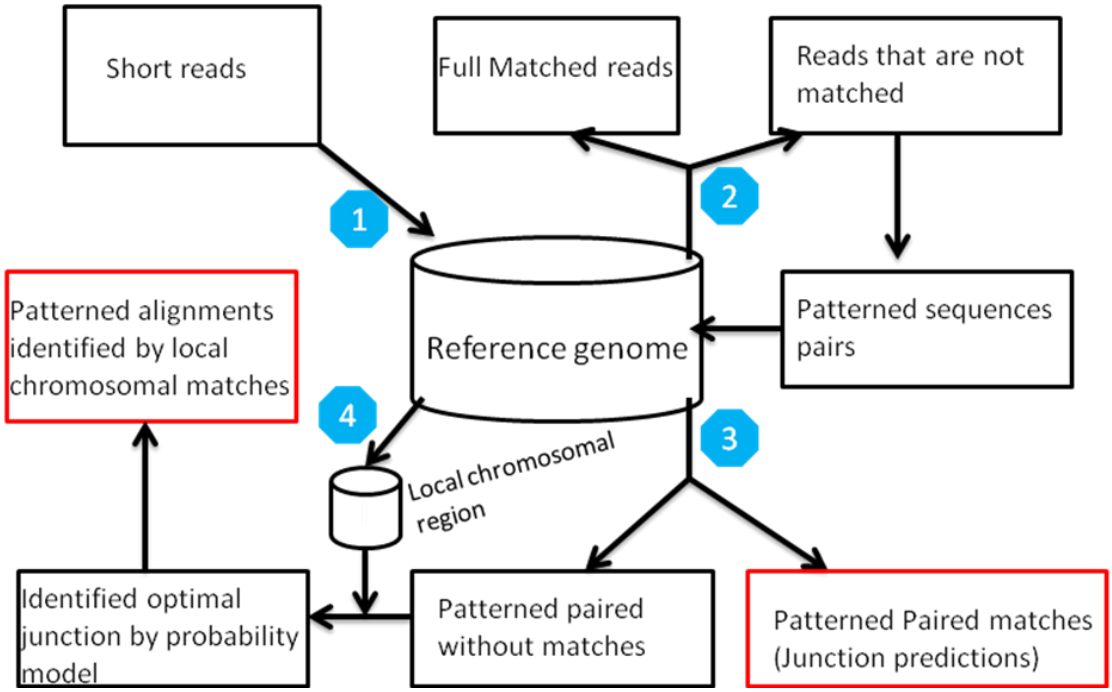
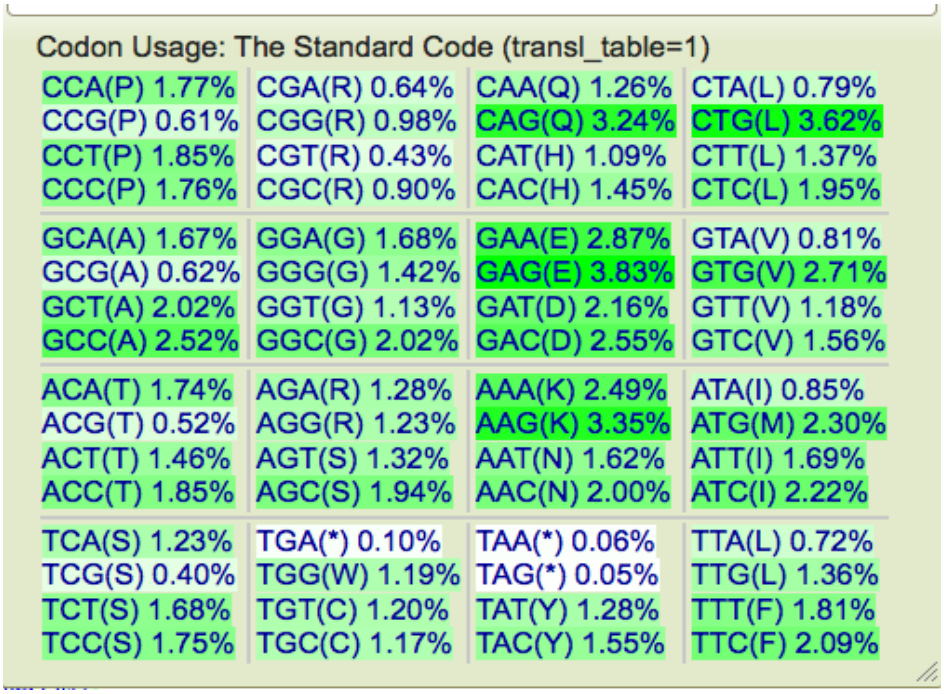


Figure 4-4. Flowchart of sequence alignment strategy from pattered fragments.

Figure 4-5. The schematic flowchart of assigning exons to genes in de novo gene reconstruction algorithm. Any exons sharing a splicing site with other transcripts are counted toward the alternative splice sites, as shown in the upper panel. No two genes share the same exon, so exons that are not able to be incorporated into current transcript will be treated as coming from a new transcript. The first and last exon from which no junctions are connecting toward outside, are extended to generate a putative incomplete exon of 50 base pairs.



Codon Usage: The Standard Code (transl_table=1)

CCA(P) 1.77%	CGA(R) 0.64%	CAA(Q) 1.26%	CTA(L) 0.79%
CCG(P) 0.61%	CGG(R) 0.98%	CAG(Q) 3.24%	CTG(L) 3.62%
CCT(P) 1.85%	CGT(R) 0.43%	CAT(H) 1.09%	CTT(L) 1.37%
CCC(P) 1.76%	CGC(R) 0.90%	CAC(H) 1.45%	CTC(L) 1.95%
GCA(A) 1.67%	GGA(G) 1.68%	GAA(E) 2.87%	GTA(V) 0.81%
GCG(A) 0.62%	GGG(G) 1.42%	GAG(E) 3.83%	GTG(V) 2.71%
GCT(A) 2.02%	GGT(G) 1.13%	GAT(D) 2.16%	GTT(V) 1.18%
GCC(A) 2.52%	GGC(G) 2.02%	GAC(D) 2.55%	GTC(V) 1.56%
ACA(T) 1.74%	AGA(R) 1.28%	AAA(K) 2.49%	ATA(I) 0.85%
ACG(T) 0.52%	AGG(R) 1.23%	AAG(K) 3.35%	ATG(M) 2.30%
ACT(T) 1.46%	AGT(S) 1.32%	AAT(N) 1.62%	ATT(I) 1.69%
ACC(T) 1.85%	AGC(S) 1.94%	AAC(N) 2.00%	ATC(I) 2.22%
TCA(S) 1.23%	TGA(*) 0.10%	TAA(*) 0.06%	TTA(L) 0.72%
TCG(S) 0.40%	TGG(W) 1.19%	TAG(*) 0.05%	TTG(L) 1.36%
TCT(S) 1.68%	TGT(C) 1.20%	TAT(Y) 1.28%	TTT(F) 1.81%
TCC(S) 1.75%	TGC(C) 1.17%	TAC(Y) 1.55%	TTC(F) 2.09%

Figure 4-6. The codon usage table in mouse chromosome 2. Each codon is represented in the nucleotide triplet. The amino acid encoded by the codon triplet is written inside the parenthesis, followed by the relative usage of codon triplet by percentage.

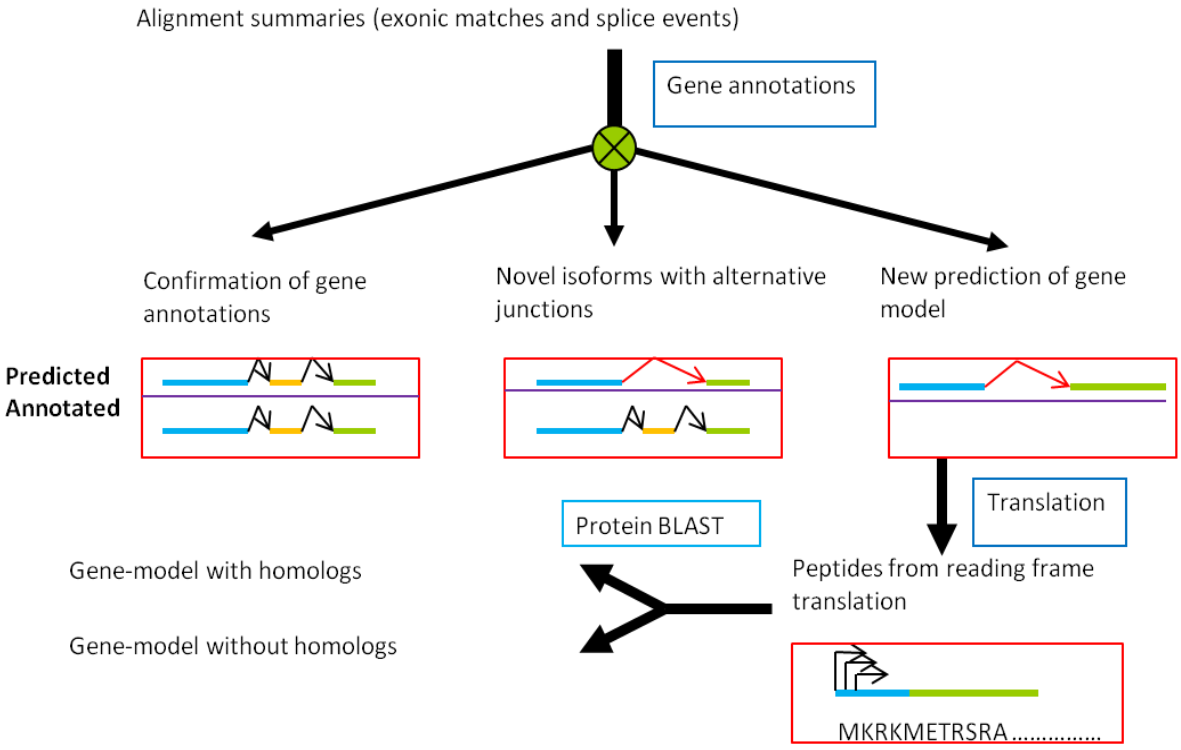


Figure 4-7. Schematic flowchart of the isoform reconstruction using PASTA on RNA-Seq datasets.

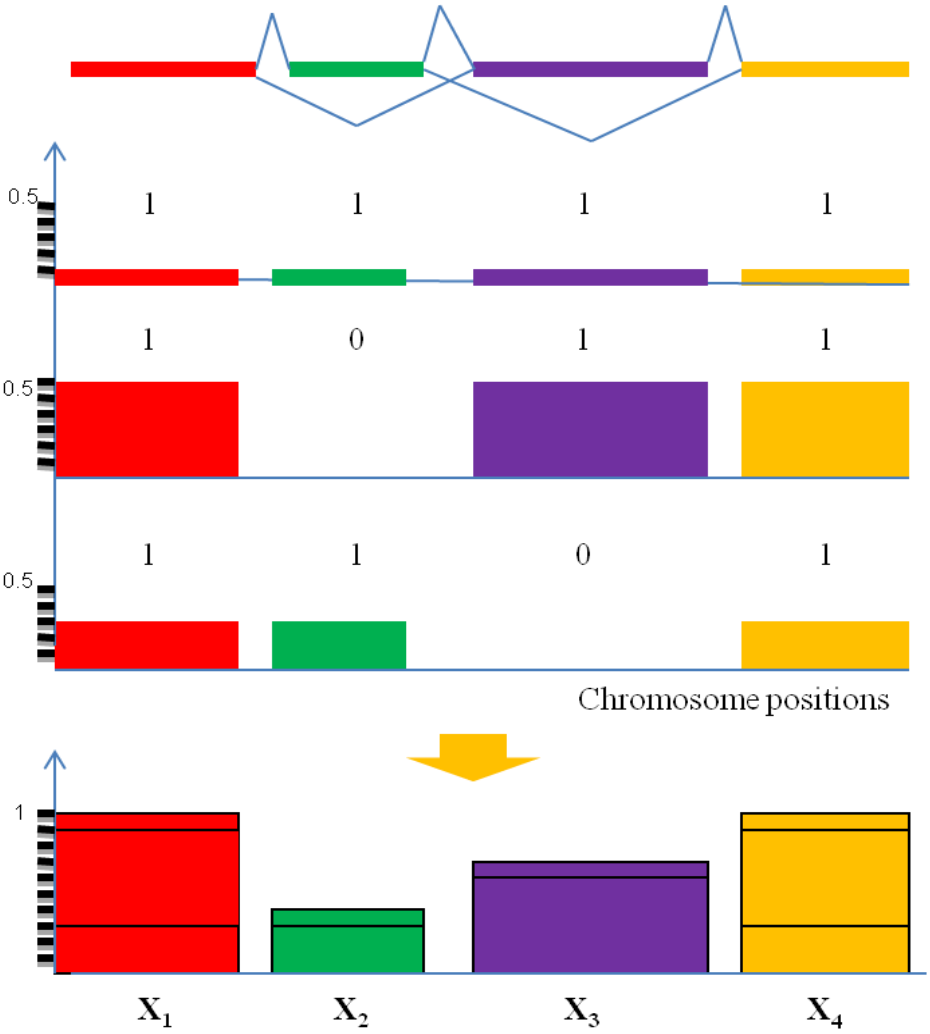


Figure 4-8. Isoform expression estimation using the Genetic Algorithm. In this figure, I display a gene with 4 exons and 3 alternative isoforms as indicated in the middle of the figure. Number 1/0 manifest a certain exon is present/absent in a particular isoforms. Genetic algorithm uses the observed average exon expression values x_1, x_2, x_3 and x_4 from the four exons respectively to infer the isoform expression level of the three alternative isoforms.

CHAPTER 5
RESULTS

I begin this chapter with a brief discussion on read mapping properties of RNA-Seq datasets using reference genome alignment. Next, I present experimental results of genome-wide splicing analysis using PASTA on mouse and maize RNA-Seq datasets. The analysis results are divided into three separate sections, and each section addresses results from one of the three modules of PASTA.

An Overview of RNA-Seq Read Distribution

I downloaded three RNA-Seq datasets from the NCBI SRA database (accession numbers “SRX000350”, “SRX000351” and “SRX000352” respectively) to study the distribution of short read alignment matches over reference genome. These datasets are aligned to mouse reference genome build NCBI37 using Bowtie. Reads mapped to reference genome with a full length match with a maximum of 2 mismatches are used compute gene expression levels. Results in Figure 5-1 and Figure 5-2 show the read mapping information on mouse chromosome 2 using RNA-Seq datasets (SRA accession number “SRX000351”). Overall, RNA-Seq short read sequence alignments using Bowtie suggest that we observe low expression abundance for the majority of known genes. Among the remaining DNA short reads that cannot be aligned to reference genome, part of them are short reads coming from two distinct exons, making it possible to predict splice junctions, a crucial step in reconstructing the gene isoforms. The first module of PASTA is designed to identify splice junctions *de novo* using a reference genome from RNA-Seq datasets.

PASTA Module 1 for Splice Junction Prediction

Intron Size Distribution from Mouse

Using gene annotations from many well-known model organisms such as mouse and human, I found that many organisms have a specific distribution of intron sizes as shown in Figure 5-3. This pattern is particularly strong for plants such as maize [17]. In general, short introns occur more often than long ones. This can be observed, for example, in existing mouse gene annotations (Figure 5-3, Figure 5-4 A).

PASTA uses the Pareto distribution to model the intron size distribution for organisms for which intron information is available. I divided the set of intron sizes into bins of 100 nucleotides and calculated the number of introns in each bin. The distribution of the intron counts normalized by the logarithm of the intron sizes in ascending order can be modeled by a Pareto distribution. I compute the Pareto score of the logarithm normalized intron i with size s_i by $f(\log(s_i) = 1 - rank(s_i)/n$ where $rank(s_i)$ is the rank of intron size **s** in the ordered list (the smallest intron has rank 1, the largest has rank n). Results show us that the logarithm normalization of the mouse intron size provides a very good approximation from the logarithm normalized Pareto distribution probability function (Figure 5-4 B).

Logistic Regression Model to Score Splice Junction Prediction

In modeling the organism specific features of splice junctions, I propose a logistic regression model based on factors such as intron size distribution, presence of canonical splice signals and short read alignment scores against reference genome, to represent the probability that a splice junction prediction is real. The logistic regression model is trained in order to obtain the optimal regression coefficients for likelihood score

computation. In obtaining the regression model for mouse, I used the publicly available RNA-Seq dataset (NCBI SRA accession number “SRX000351”) and RefSeq gene annotations. Initial RNA-Seq short reads were aligned to reference genome with PASTA to produce potential junctions. Subsequently, I analyzed each predicted junction to compute its Pareto score, check for the presence of canonical splicing signal and branch point sequence, and record its sequence alignment identity. Finally, we compared the junction position with RefSeq annotations to determine whether it is a real junction or not. These values are then fit in a logistic regression model to obtain the final regression coefficients. Overall, logistic regression enables PASTA to compute the likelihood score for every new junction predicted, thus providing a measure of the quality of splice junctions. Finally, I generated a residue plot (Figure 5-5) of the likelihood score by subtracting the predicted score from the real score (a score of 1 is used for a real junction and -1 for false predictions on the basis of RefSeq gene annotations). Result shows us that the regression model is able to identify a majority of true junctions with very high sensitivity, and only introduces a small number of false positives. By setting the score threshold of 0.5, PASTA is able to correctly predict 99% true splice junctions and less than 5% false positives.

Comparison of PASTA and Tophat on Splice Junction Prediction with Single End Simulated RNA-Seq Data

As a first test of the performance of PASTA, I have compared its ability to detect known splice junctions against Tophat, currently one of the most widely used programs for splice junction detection from simulated RNA-Seq data. I have generated several simulated datasets of RNA-Seq short reads, corresponding to different sequencing depths, and I measured the sensitivity and specificity of each program. The average

read depth (the average number of times each transcript's nucleotide is sequenced) in the four simulated single-end RNA-Seq datasets is 1, 2, 4 and 8 respectively. For example, given 10 million single end reads of 50 base pairs, a read depth of 1 is approximately equals to 2 RPKM [54] I simulated single-end RNA-Seq reads with a length of 50 base pairs from known mouse transcripts in RefSeq annotations introducing random sequencing errors at a frequency of 1/1000 base pairs and single-nucleotide mutations at a frequency of 5/1000 base pairs. PASTA shows very high sensitivity and specificity in splice junction detection as shown in the ROC curve at Figure 5-6. In general, as read depth increases, sensitivity increases (since detecting junctions becomes easier) but specificity also decreases (since the number of false positives increases with the number of reads). The results show that PASTA consistently displays higher sensitivity than Tophat (on average 20% to 40% higher), especially at low coverage levels, and a similar specificity (Figure 5-7). PASTA is therefore well-suited for identifying "rare" splicing events, reducing the risk of missing splicing events critical for AS analysis. This also indicates that the use of PASTA may lead to a reduction in sequencing costs, for example by multiplexing more samples in the same run, since it is able to produce reliable results even at low sequencing depths.

Comparison of PASTA and Other Pipelines on Splice Junction Prediction using Paired-end Simulated RNA-Seq Data

I also compared PASTA with several other splice junction prediction pipelines using simulated paired-end reads of 100 nucleotides from mouse transcripts generated in a previous study [78]. Two separate simulated scenarios with different polymorphism and error rates are simulated in these datasets. In both simulated datasets, PASTA

performed much better than MAPSPLICE, SPLICEMAP or Tophat, and PASTA also achieved performance that is comparable to GSNAP or RUM (Figure 5-8).

Overall, simulations with various sequencing depths show that PASTA is very sensitive in identifying splice junctions even when the sequencing depth is relatively low. PASTA could effectively identify splice junctions with reads of small to medium size (30 to 70 base pairs) and rare transcripts for which patterned alignment is specifically designed. Even with paired-end reads with long read size and high sequencing depth as described in the second simulation, PASTA still achieves performance comparable to the best software.

Comparison of PASTA and Tophat on Maize and Human RNA-Seq Datasets from Illumina and SOLiD Platform

I have further compared PASTA with Tophat on their splice junction identification performance with a variety of samples, including RNA-Seq datasets from maize and human from different sequencing platforms. The comparisons of PASTA and Tophat with respect to splice junction prediction accuracy are shown in Figure 5-9. Results show that PASTA is able to detect the majority (80%~98% of the total) of splice junctions identified by Tophat. In addition, PASTA finds a large number of additional splice junctions not reported by Tophat, which in turn proves PASTA's ability to identify transcripts even in low abundance. Overall, the performance of PASTA is consistent using datasets from different organisms on distinct RNA-Seq platforms, and this study enables us to use PASTA in a wide range of applications, including datasets from different organisms generated by various RNA-Seq platforms.

Validation of PASTA Junction Predictions Using RNA-Seq Datasets from Mouse Mbnl3 Isoform Knockout Samples

I have performed three RNA-Sequencing runs in order to generate initial validation data for PASTA. In the following I provide a description of the experiments and I present our preliminary analysis of the results.

In order to evaluate the performance of PASTA in detecting splice junctions, I ran Tophat on the *Mbnl3* isoform knockout data using default options, and I compared the number of known junctions (from the RefSeq database) identified by the two programs. In the Tophat configuration, I configured the option 'mate-inner-dist' as 250 and quality score version as 'solexa1.3'. Default values are used for other options. Since I do not have an independent way of confirming the predicted junctions, I have used the set of junctions derived from the validated dated gene models in ENSEMBL as the "golden standard" against which to measure performance. The ENSEMBL dataset contains more experimentally validated genes from different sources. Results in Table 5-1, 5-2 and 5-3 contain results of the comparisons between PASTA and Tophat on *Mbnl3* data from two experimental runs: Run1, which is performed on February 2011, and another experiment Run2, which is performed on December 2010.

The results show, first of all, that PASTA detects a higher number of junctions than Tophat, especially with shorter read lengths, and the difference increases with the number of reads. The second part of each table focuses on known junctions (from RefSeq isoforms). I calculated the number of ENSEMBL junctions identified by both programs, and expressed this number as a percentage of the ones identified by Tophat. The resulting value, that indicates the percentage of Tophat junctions that PASTA correctly identifies, is consistently at 95% or higher in Run 1, and at 97% or higher in

Run 2. It is interesting to note that, in Run 2, higher percentages are somewhat correlated with lower numbers of reads. Additionally, PASTA identifies an extra 70% or higher ENSEMBL junctions in Run 1, and 25% or higher ENSEMBL junctions in Run 2 that are completely absent in Tophat predictions. Some of these extra ENSEMBL junctions identified by PASTA map to exons that are sequenced at low levels or are supported by very few junction-spanning reads; these reads are therefore crucial in reconstructing correct gene structures.

I also computed the coverage of PASTA predicted junctions (the number of reads supporting a specific junction), and analyzed how it relates to model probability scores and the ratio of canonical junctions. Table 5-4 displays the average probability score and the frequency of canonical junctions as a function of the coverage level (expressed as reads/junctions) on junctions identified by PASTA. The results show that that there is a significant correlation of read coverage and the probability score, and the higher coverage is usually a strong indicator of real junctions having high probability scores and appearing in ENSEMBL junctions. These results are further supported by Table 5-5. Using ENSEMBL known genes, I can see that PASTA predicted junctions that appear in known genes exhibit higher average probability scores and higher coverage. In addition, Table 5-6 shows that canonical junctions, that are frequently accompanied by high probability score and high coverage, are significantly different from non-canonical junctions in average scores. Finally, in order to see the average coverage for PASTA predicted junctions that appeared in ENSEMBL know genes, I compute the number of total number of reads that falls on these junctions. These results suggest that PASTA predicted junctions that appear in ENSEMBL genes are mostly canonical junctions

(97% or more), and are normally supported by high probability score and high coverage as shown by Table 5-4.

This demonstrates that PASTA is more sensitive than Tophat in identifying known junctions, in particular when the sequencing depth is low. The third part of each table reports the number of additional junctions identified by both programs. Although it is not possible to know at this stage whether these junctions are real or is false positives, a substantial fraction of them are identified by both programs, especially with the longer read length and higher sequencing depths.

Analysis of PASTA splice junctions appearing in gene annotations

In order to see the average coverage of splice junctions appearing in ENSEMBL mouse gene annotations, I computed the number of reads supporting each junction predictions. Results show that a very large portion of these splice junctions are supported by multiple short reads (Figure 5-10 and Table 5-5). This result is observed across different biological replicates, providing PASTA's capability to identify splice junction in the context of low abundance to high expression levels.

Analysis of splice junctions uniquely identified by PASTA and not by Tophat

Previous studies showed that PASTA is able to generate more splice junctions that are uniquely mapped to gene annotations compared with Tophat. In this section, I conduct a detailed analysis of these PASTA specific ENSEMBL junctions, and the results are summarized in the Table 5-4. The results show that almost half of these junctions are supported by two or more reads in both experiments, and 15% to 20% of the short reads are supported by five or more reads, which indicates that the majority of these PASTA specific annotated junctions are appearing in high abundance but are absent in Tophat predictions.

Analysis of Putative New Junctions

RNA-Seq enables in-depth analysis of putative new junctions. In this study, I used PASTA and Tophat to identify putative new junctions which do not appear in existing ENSEMBL annotations using the Mbnl3 isoform knockout RNA-Seq datasets. Results are summarized in Table 5-3. PASTA in general produces more putative new junctions, and the overlap between PASTA and Tophat predictions is relative small compared to the total number of ENSEMBL known junctions. These putative new junctions fall on two categories: alternative sites and exon skipping (Figure 5-12), and they enable scientists to discover new alternative isoforms and/or novel transcripts.

Finally, in order to evaluate the PASTA predictions, I selected a list of splice junction candidates for bench validations using RT-PCR experiments on Mbnl3 RNA-Seq samples. The list of candidate genes is shown in Table 5-20.

Alternative Splicing Analysis using RNA-Seq data from Maize Samples

In order to test the applicability of PASTA to genomes with different characteristics, we analyzed a dataset from maize RNA-Seq experiments, consisting of three samples, the "FSU" control sample, and the "sample3" and "sample4" mutant samples. PASTA identified more than 50% of splice junctions appearing in maize genome annotations, and it also discovered novel transcripts. Among them, alternative 5’ or 3’ splice sites of existing junctions occur most frequently in maize.

Profiling of alternative 5’ or 3’ sites from maize RNA-Seq datasets

Previous studies showed that genes in plants such as maize usually contain more 5' or 3' alternative sites compared with higher organisms such as mouse and human [18]. In our analysis, I found a total of 1900 and 1262 alternative splicing sites from sample3 and sample4 respectively. A comparison between the two datasets

produced 552 novel alternative splice sites present in both datasets. In contrast, 4324 alternative sites are identified on the maize control sample. Overall, a total of 312 common alternative splice sites are observed in all three datasets of the maize RNA-Seq datasets (Table 5-7).

Our analysis showed that 5' or 3' alternative splice sites are found in a high frequency using three RNA-Seq datasets from maize samples. The introduction of alternative splice sites may cause a disruption of the original reading frame for coding transcripts. Therefore, to characterize the effects of alternative splice site usage, I translate the putative novel isoforms into candidate peptides for homologous sequence analysis to see whether the introduced alternative sites will disrupt the reading frame of the original coding sequences. Our results show that 117 out of 438 reported alternative splice sites lead to a removal a multiple of 3 nucleotides, thus preserving the reading frame. This result also shows that there is no strong preference in preserving the original ORF in the choices of alternative splice sites, which may contradict previous finding that alternative splicing events tend to preserve the reading frame [79]

Global cassette exon (exon-skip) profiling using maize RNA-Seq datasets

PASTA analysis shows that the number of exon skipping events discovered using the RNA-Seq platforms Illumina and SOLiD Applied Biosystems are very different from each other. In general, I identified three times more cassette exons with the Illumina platform compared with the SOLiD Applied Biosystems. The difference may be a result of the large number of reads and a larger read size in Illumina platform than the SOLiD RNA-Seq datasets. A comparison between the two datasets in the SOLiD platform produced a total of 83 common exon skips. Overall, 58 exon skips were identified on all three samples shown in Table 5-8. These exon skips are appearing in

potential new isoforms, and the characterization of these new isoforms may help scientists discover novel gene regulatory mechanisms.

Identification of putative novel junctions using maize RNA-Seq datasets

PASTA analysis using maize RNA-Seq datasets shows that approximately 10% of the predicted splice junctions do not appear in maize gene annotations (Figure 5-13), but the prediction drops significantly in the Tophat analysis, which shows an average of approximately 2000 novel splice junctions using the three RNA-Seq datasets from maize. Overall, these new splice junctions may come from putative new genes or pre-mRNAs of micro-RNAs and/or other regulatory RNAs, thus providing invaluable information in studying alternative splicing and novel gene assembly.

Analysis of splicing signal using maize RNA-Seq dataset

In order to study the usage of canonical splice signals using maize RNA-Seq datasets, I computed the number of canonical and non-canonical splice signals predictions (Figure 5-14). Results show that approximately 95% to 97% of the splice signals are canonical, which confirms the widespread usage of canonical signals in maize. However, minor splice signals such as GC-AG are also observed.

Visualization of PASTA Predicted Junctions

PASTA provides a graphical representation of the expression of exons and splice junctions from the output files. Expression levels from RNA-Seq experiment are displayed in parallel with reference gene annotations to allow for the observation of constitutive and alternative spliced exons. For example, using mouse RNA-Seq dataset Run 1, Figure 5-11 shows that mouse gene *Dgcr6* is alternatively spliced with the first and second exon replaced by an alternative new exon. In addition, The PASTA prediction output file is in the standard BED or BedGraph format. Therefore, the output

file can be uploaded to the UCSC genome browser: http://genome.ucsc.edu/ to allow straightforward visualization in the genome browser.

PASTA Module 2 for De Novo Isoform Reconstruction

De Novo Gene Discovery from RNA-Seq Datasets using Maize Samples

Studies using maize RNA-Seq datasets showed that PASTA is able to identify transcripts not annotated in existing maize genome annotations. Overall, approximately 80% of the reconstructed transcripts from maize RNA-Seq sample-3 and sample-4 fall on existing gene annotations, and 14% of the predictions are alternative isoforms of the existing genes and the remaining 6% of the predictions are new genes that do not appear in maize gene annotations (Figure 5-15). I performed isoform reconstruction using all three maize RNA-Seq datasets and the results are shown in Figure 5-16.

De Novo Gene Discovery from RNA-Seq Datasets using Mouse Samples

I used PASTA to reconstruct transcripts from the RNA-Seq datasets of mouse Mbnl3 Isoform knockout experiments. Two RNA-Seq samples of 49 and 19 million paired-end reads with the read size of 40 base pairs were analyzed in this study. Results in Figure 5-17 show us that the majority of predicted isoforms are found in existing RefSeq annotation. I tested PASTA and Cufflinks on both datasets to compare their ability to reconstruct mouse transcripts. For Cufflinks analysis, RNA-Seq reads were initially aligned with Tophat using default options and ‘mate-inner-dist’ of 250 to collect mapping results in BAM format as the input file for subsequent Cufflinks assembly. Cufflink default options are used during the de novo gene assembly. Reference gene annotations were obtained from ENSEMBL.

The results shown in Figure 5-18 indicate that the number of predicted genes with a perfect identity to annotations is still relatively a small portion because the RNA-

Seq datasets contain a small number of reads. Therefore, I defined the matching identity of *de novo* prediction as the ratio of predicted exons that are identical to the ones in existing gene annotations for each transcript. Results show that PASTA is able to assemble around half of the UCSC genes with at least 50% identity or greater, while Cufflinks showed a much lower prediction identity.

The results in Table 5-9 demonstrate that Cufflinks identifies 2 times more genes than annotations using RNA-Seq datasets in Run1, but Cufflinks exons are approximately 2 to 3 times longer than the average length of exons recorded from annotations. On the other hand, the average number of exons in each predicted isoform is only 1/5 of that found in the annotations. In addition, one of the key goals of this work, namely identifying alternative splicing isoforms, is missed in Cufflinks predictions, as *de novo* gene models from both RNA-Seq datasets showed a very low average number of isoforms per gene. These results showed that Cufflinks is not able to effectively reconstruct transcripts that are typically composed of multiple exons with alternative splicing isoforms in the mouse transcriptome. In contrast, PASTA produced a gene annotation set that is comparable to existing annotations. The average number of exons from PASTA's predicted genes is 7, which is much closer to the value in ENSEMBL gene annotations. From Table 5-9, I can also see that both PASTA and Cufflinks predict a shorter average isoform length compared with annotations. Because PASTA uses coding transcripts in this analysis, the typically long Un-Translated Regions (UTR) are not counted towards the isoform length. On the other hand, PASTA also trims the length of first exon and last exon if there is not enough sequencing coverage to accurately identify the beginning of the first exon and the end of the last exon.

These two experiments demonstrate that PASTA's performance is consistent for RNA-Seq experiments with very different read throughput size. Nevertheless, a reduced number of genes is predicted when read throughput is low (19 million reads of 40 base pairs). PASTA is still able to find genes closer to annotations in terms of features such as the average number of exons per transcripts. PASTA is particularly optimized for coding transcriptome assembly. The inclusion of coding utilization frequency enhances the accuracy of coding transcript assembly these transcripts have with high matching identity with existing genome annotations. In this study, coding utilization frequency from only a portion of existing mouse genes leads to effectively reconstructing coding exons that show high similarity with existing gene annotation.

PASTA Module 3 for Differential Expression and Alternative Splicing Analysis

An Overview of Mbnl3 RNA-Seq Read Alignments

I have conducted a genome wide analysis of gene expression on RNA-Seq datasets from *Mbnl3* isoform knockout experiments, using PASTA, MISO [40] and IsoEM [80]. In total, three RNA-Seq experiments are performed on Illumina Genome Analyzer II. Run1 and Run2 come from paired-end sequencing on forelimb tissue of embryonic stage E14.5 *Mbnl3* isoform knockout mouse of 40 base pairs of read size and 51 base pairs respectively. Paired-end RNA-Seq datasets from Run3 come from an Mbnl2 isoform knockout experiment and have a read size of 40 base pairs. Alignment summaries with bowtie and PASTA are shown in Table 5-1. The *Mbnl3* isoform knockout is confirmed by PASTA as shown in Figure 5-22 and Figure 5-23. From the figures, I can see the absent expression level of skipped exon 2 (the one before the last exon because *Mbnl3* is transcribed on the reverse strand).

In order to check the consistency of expression for the *Mbnl3* RNA-Seq datasets, I have computed the differences of individual exon expression levels by subtracting the normalized expression level, for all RefSeq annotated exons. The results are displayed in Table 5-11.

The normalized exon expression level x is calculated as:

$$x = \left(\sum_{i=1}^{n} c_i \right) / (n * d) \tag{5-1}$$

In this formulation, n is the number of nucleotides in the exon, and c_i is the number of RNA-Seq short reads that cover nucleotide position i in the exon under study. The results are normalized by dividing the total RNA-Seq throughput of the lane in millions (denoted with d in the formula), for example, $d = 10$ for a throughput of 10 million reads. Results in Table 5-11 show that the gene expression levels of the majority of the genes are consistent between samples of the same RNA-Seq datasets. Therefore, the normalization of the RNA-Seq throughput by read size is adequate for subsequent analysis.

Differential Expression of Genes from RNA-Seq Datasets of Mbnl3 Isoform Knockout Experiments Using MISO and IsoEM and PASTA

I used MISO to compute genes with differentially expressed isoforms. Default options are used for the MISO computation. MISO predictions are measured by the Bayes factor, which shows the likelihood that a gene is differentially expressed. The Bayes factor is calculated for every control and mutant sample, and thus I have computed all pairwise comparisons between each control and mutant sample. Results in Figure 5-24 show the number of MISO predictions using *Mbnl3* RNA-Seq datasets on Run1 with a Bayes factor of 10. Results show 1584 genes are differentially expressed

between control sample 1 and mutant sample 7. However, two of the pairwise comparisons failed because of MISO internal errors. For the correctness of our results, I have removed the incomplete pairwise comparisons between sample 1 and 5, and sample 3 and 5 from our analysis. Overall, approximately 1200 to 1600 genes are differentially expressed in the remaining pairwise comparisons. Among predictions from each pairwise comparison, I found 375 genes are differentially expressed in every pairwise comparison from MISO. I also repeat the analysis form Mbnl3 RNA-Seq dataset on Run2. Results in Table 5-10 show us that 2449 genes are differentially expressed in Run1 and 2783 genes are differentially expressed in Run2.

I further performed this analysis using the IsoEM algorithm, which computes individual isoform expression levels using the Expectation Maximization (EM) algorithm on gene and isoform annotations from ENSEMBL. IsoEM generates read counts for each isoforms, and uses these read counts to compute the fold changes for every isoform between every control and mutant sample, as the ratio between the mean expression level in control and mutant samples. Using IsoEM we obtained 5578 differentially expressed genes in the Run1 and 12207 differentially expressed genes in Run2, using a threshold on the mean expression fold change of 2. However, among these differentially expressed isoforms, only 226 of them appear in both RNA-Seq experiments Run1 and Run2. Subsequently, I test the overlap between IsoEM and MISO in the predictions of differentially expressed isoforms. I found a total of 27 isoforms that are differentially expressed by both methods in Run1 and 35 isoforms that are differentially expressed by both methods in Run2 (Table 5-12).

Finally, I performed the same analysis using PASTA with fold changes calculated from the average expression level of mutant samples divided by control samples. PASTA identified a total of 493 and 233 genes with differentially expressed isoforms by a fold change of 2 in Run1 and Run2 respectively. Overall, I have seen that PASTA observes a smaller number of differentially expressed transcripts, and this variability could be explained by the following two facts. MISO and IsoEM perform pairwise comparisons between each control and mutant sample at a time, which produces more candidate genes with differentially expressed isoforms. In contrast, PASTA works on the control samples and mutant samples as two populations, performing a single comparison. In addition, PASTA requires stronger consistent expression levels (smaller variance and variability) among control samples and mutant samples in the prediction.

Overall, analysis from MISO, IsoEM and PASTA show a large difference in the number of gene predictions with differentially expressed isoforms, and the overlaps between the predictions is also small due to the distinct behaviors of different software. Therefore, scientists should consider different methods in selecting the targets when performing analysis of differential expression. Table 5-13 displays top 10 differentially expressed genes which are detected by all three software, from both Run1 and Run2 of the Mbnl3 datasets.

Isoform Expression Level Estimation on Mbnl3 RNA-Seq Datasets using PASTA's Genetic Algorithm

I used a Genetic algorithm to compute isoform expression levels for genes having multiple transcripts using RNA-Seq datasets from Run1 and Run2. I then summed the estimated isoform expression levels from the Genetic algorithm in control samples and mutant samples respectively. Finally I obtain the fold change of average

expression in mutant divided by controls samples. This approach enables us to select genes with differentially expressed isoforms, such as an up-regulation of expression in the first isoform and down-regulation of expression in the second isoform. Results in Table 5-14 shows top candidate genes that are differentially regulated between control and mutant samples. For example, one isoform (RefSeq mRNA accession ID NM_019923) is 3.41-fold up-regulated in the mutant samples while another isoform (mRNA accession ID NM_010586) is down-regulated by 0.55 fold in mutant samples in mouse gene Itpr2.

Analysis of Genes with Differentially Expressed Isoforms on RNA-Seq Datasets from Mbnl2 Isoform Knockout Mice using PASTA and Other Software

Overall alignment summaries for Mbnl2 RNA-Seq datasets

This section describes the results of the analysis of RNA-Seq datasets from Mbnl2 isoform knockout experiments. The RNA-Seq experiments were performed on Illumina Genome Analyzer II platform with paired-end RNA-Seq datasets from Mbnl2 isoform knockout experiment. The read size is 40 base pairs. The sample sizes and their alignment summaries with Bowtie and PASTA are shown in Table 5-15. By default, PASTA generate a large number possible junction clusters, and Table 5-16 shows the fraction of canonical and non-canonical junctions. Finally, PASTA uses the scoring filter from the regression model to generate junction candidates by retaining good quality alignments and high likelihood score for subsequent analysis.

Differential expression of Mbnl2 isoform knockout using MISO and IsoEM and PASTA

In the analysis of RNA-Seq datasets from Mbnl2 isoform knockout mice, I used the similar approaches to the ones described for Mbnl3 RNA-Seq datasets. First, I computed the fold change of exon and junction expressions, and then the isoform

expression levels are estimated by the PASTA's Genetic algorithm. Second, I analyzed the datasets using MISO and IsoEM. The results are shown in Table 5-17. Overall, MISO identified approximately 1500 genes with differentially expressed isoforms, using a Bayes factor of 10 and a fold change of 1.5. On the other hand, IsoEM predicted a much larger gene list using the same fold change threshold, and PASTA identified a number of genes between MISO and IsoEM. Finally, I used the Student t test and K-S test to find the genes with statistically different mean values in exon and junction expression. The results in Table 5-19 show that 80% of genes identified by PASTA by more than1.5 fold change in gene expression level, or through the K-S test are validated by exon junction microarray experiments. The t-test identifies fewer targets probably due to the underlying assumption of normality. Finally, I only identify less than 1/5 of the target genes using PASTA's Genetic algorithm, which requires the gene under studied to have strictly equal or greater than two isoforms in the existing gene annotations.

Gene ontology analysis for Mbnl2 candidate genes with differentially expressed isoforms

I am interested in analyzing Gene Ontology information about the 81 genes containing differentially expressed isoforms according to MISO, IsoEM and PASTA in order to extract information about the molecular functions affected by Mbnl2 isoform knockout experiment. In general, I found that 30 gene ontology categories are affected as shown in Table 5-18.

Graphical Representation of Gene Expressions from Control/Experimental Groups

Finally, PASTA module 3 provides a user-friendly graphical representation of gene expression using average expression estimates from individual exons and splice junctions. The figures are generated on demand to allow users to preferentially display

genes of interests, as in Figure 5-19. In addition, PASTA provides users with an option to produce figures with normalized exon and intron sizes to make the display more readable, especially for genes with very long introns and relatively short exons (see Figure 5-20 and Figure 5-21). PASTA enables user to choose from three different representations of the gene expression profiles.

PCR Assays

Gene Targets and PCR Reactions

In order to validate PASTA's ability to detect splice junctions, minor splice sites, and differential isoform expression, I select a total of 23 splicing junctions from the Mbnl2 datasets and tested them experimentally. 5 of these candidate targets contained the minor splice site AT-AC, 4 candidate targets contain GC-AG, and the remaining 14 candidate targets contain exon skips or mutually exclusive isoforms. Among these 14 candidates, 5 of them may contain differentially expressed isoforms, as indicated by analysis using PASTA module 3 that shows a statistically significant different expression fold change for isoforms between control and knockout samples. Primers were designed using NCBI Primer-BLAST tool and were ordered from Integrated DNA Technologies (Table 5-20 and Table 5-21), Inc. Each primer contains an approximately 25nM initial yield. I diluted each primer to 50 µM by adding 500 µL ddH_20, and then I further diluted each primer into 5 µM by adding 10µL forward and 10µL reverse primers with 80 µL ddH_20 to make 100 µL primer mixtures for each primer pair. I used 12 µL cDNA of embryonic stage 14.5 forebrain samples from control and Mbnl2 knockout mice respectively. These cDNA templates from control and Mbnl2 knockout samples were diluted into 24 µL.

The PCR reaction was composed of the following: 1µL cDNA from the template, 2 µL primer mixtures for each gene, 2 µL of 10X buffer and 2 µL of 2.5mM dNTP, 0.2 µL of Taq enzyme and 12.8 µL of ddH20. The PCR thermal cycling parameters were as follows: 95°C for 30 seconds, 35 cycles of 95°C for 30 seconds for denaturing step and 60°C for 30 second for product annealing and another 68°C for product extension, and 68°C for 5 minutes for final extension.

PCR Reaction Results

Results show that all splice junctions with minor splice sites are detected by gel electrophoresis (Figure 5-26). In addition, 13 of the 14 splice junction targets with exon skip or mutually exclusive exons were validated using PCR reactions and gel electrophoresis (Figure 5-25), and I only failed to detect one exon skipping event which is possibly caused by extremely low gene expression level. In general, PASTA is very sensitive in identifying genes with differentially expressed isoforms. For example, in mouse gene Slain2, the exon-skipping isoform is up-regulated in knockout samples, as indicated by decreased skipped exon expression level and a large number of junctions with exon skip. This change is shown very clearly in the gel electrophoresis results in Figure 5-26. Another gene Csnk1d show very similar regulation pattern where exon-skipping isoform is substantially up-regulated in the knockout sample.

The PASTA Software

Website

PASTA is available for download in http://genome.ufl.edu/rivalab/PASTA

Installing PASTA

After downloading the package, copy it to an appropriate directory and execute the following command: tar -xvf PASTA-1.0.tar.gz, which will create a directory called pasta-1.0/ with executable for stand-alone applications.

The PASTA Splice Junction Aligner

PASTA takes as input a file containing RNA-Seq reads in FASTQ format. The initial alignment is performed using the bowtie program that must therefore be installed and accessible in order for PASTA to work. Bowtie indexes can be downloaded from the bowtie website or created using the bowtie program. In addition, a reference genome must be provided to PASTA, and it must be the same reference genome used for bowtie indexing. The splice junction prediction program is based on a logistic regression model, whose coefficients are provided in a configuration file. The model coefficients are pre-computed and are different for each organism. The coefficients provided with the current version of PASTA are for mouse; coefficient files for other organisms will be made available in the near future. PASTA generates a file containing the location of predicted splice junctions and an alignment summary containing a list of read alignments in SAM format.

The PASTA *De Novo* Isoform Reconstruction

The PASTA isoform reconstruction module requires a file containing splice junctions and a short read mapping file from Bowtie or Tophat. In addition, if the codon utilization frequency table is available, user could utilize the table to optimize gene predictions. Finally, PASTA generates a file gene model prediction file containing the reconstructed genes in GTF file. In addition, a separate file containing the peptide sequences is generated using the optimized ORF for each isoform prediction.

Prerequisites

PASTA is a GNU/Linux command-line 64bit executable. It relies on the presence of the bowtie program if original RNA-Seq data in FASTQ format is used as input. Therefore, the following executables available: bowtie (required), bowtie-inspect (recommended) and bowtie-build (recommended).

Reference Genomes and Bowtie Indexes

PASTA requires a complete reference genome for RNA-Seq read mapping and local sequence alignments. Reference genome sequences must be placed in a directory accessible to PASTA. In addition, reference genome needs to be separated into chromosomes files in FASTA format, with the identical name in chromosome file prefix and FASTA header, such as “chr1”. Finally, PASTA requires Bowtie indexes to be generated using the same reference genome, and they are placed in a directory accessible by PASTA.

Table 5-1. Number of reads and junctions detected using Mbnl3 RNA-Seq datasets. This table displays the total number of reads, the total number of junctions identified by both programs, and the ratio between these two numbers for Run 1 and Run2 respectively.

Run			Number of reads (millions)	PASTA Junctions	TopHat junctions	Ratio
1	Control	Lane 1	19.2	165541	80211	2.064
		Lane 3	15.4	149797	72908	1.828
		Total	34.6	195731	112581	1.739
	Mutant	Lane 1	21.8	169493	72908	2.325
		Lane 2	17.9	157481	82036	1.920
		Lane 3	22.3	162408	81823	1.985
		Lane 4	39.2	202157	59014	3.426
		Total	101.2	287568	152196	1.889
2	Control	Lane 1	29.9	166050	140831	1.179
		Lane 2	8.74	141885	107399	1.321
		Lane 3	10.2	144879	110459	1.312
		Total	48.84	210016	157949	1.330
	Mutant	Lane 1	27.6	148238	113908	1.301
		Lane 2	10.6	160885	124606	1.291
		Lane 3	25.4	175240	133601	1.312
		Lane 4	25.6	177388	133539	1.328
		Total	89.2	250991	167664	1.497

Table 5-2. Number of junctions detected which appears in ENSEMBL known genes using Mbnl3 RNA-Seq datasets. This table displays the number of junctions in ENSEMBL known gene models identified by PASTA and TopHat for Run 1 and Run2 respectively.

Run			PASTA	TopHat	Common	Common / PASTA	Common / TopHat
1	Control	Lane 1	128811	65117	63063	0.490	0.968
		Lane 3	120465	67552	65252	0.542	0.966
		Total	140083	86148	83674	0.597	0.971
	Mutant	Lane 1	129099	57615	55770	0.432	0.968
		Lane 2	122237	67038	64517	0.528	0.962
		Lane 3	123860	65568	63078	0.509	0.962
		Lane 4	142097	41084	39695	0.279	0.966
		Total	163462	98757	95854	0.586	0.971
2	Control	Lane 1	130899	115751	111098	0.849	0.960
		Lane 2	119397	94638	91672	0.768	0.969
		Lane 3	119950	96743	93614	0.780	0.968
		Total	146117	123247	118544	0.811	0.962
	Mutant	Lane 1	122889	99377	96287	0.784	0.969
		Lane 2	127854	106340	102840	0.804	0.967
		Lane 3	132544	111252	107418	0.810	0.966
		Lane 4	134049	111334	107571	0.802	0.966
		Total	156339	126177	121633	0.778	0.964

Table 5-3. Number of junctions not appearing in ENSEMBL known genes using Mbnl3 RNA-Seq datasets. The third part displays the number of additional junctions (not in ENSEMBL annotation) identified by PASTA and TopHat for Run 1 and Run2 respectively.

Run	Group		PASTA	TopHat	Common	Common / PASTA	Common / TopHat
1	Control	Lane 1	36702	15094	3267	0.089	0.216
		Lane 3	29331	14416	3098	0.106	0.215
		Total	55647	26433	5368	0.096	0.203
	Mutant	Lane 1	40393	15293	2589	0.064	0.169
		Lane 2	35243	14997	2990	0.085	0.199
		Lane 3	38547	16255	3110	0.081	0.191
		Lane 4	60059	17930	2234	0.037	0.125
		Total	124104	53439	7947	0.064	0.149
2	Control	Lane 1	35150	25080	10251	0.292	0.409
		Lane 2	22487	12716	4785	0.213	0.376
		Lane 3	24928	13716	5190	0.208	0.378
		Total	63898	34702	15181	0.238	0.437
	Mutant	Lane 1	25348	14531	5710	0.225	0.393
		Lane 2	33030	18266	7254	0.220	0.397
		Lane 3	42695	22349	9368	0.219	0.419
		Lane 4	43338	22205	9290	0.214	0.418
		Total	94651	41487	19214	0.203	0.463

Table 5-4. Average probability scores and percentages of canonical junctions by junction coverage using Mbnl3 RNA-Seq datasets. Only junctions with a maximum coverage of 100 are studied. A cut off value is given for each analysis with Run1 and Run2. The data in the left side contains results having coverage smaller or equal to the given cut off value, and the data in the right side contains results having coverage greater than the given cut off value.

Cut off	Run	Lane id	Junctions	Avg Score	Percent Canonical	Junctions	Avg Score	Canonical Fraction
1	1	1	182951	0.208	0.169	161315	0.71	0.706
		3	151512	0.215	0.185	146821	0.71	0.72
		5	225167	0.181	0.143	168571	0.697	0.683
		6	214297	0.167	0.139	160157	0.665	0.673
		7	196135	0.176	0.141	186288	0.607	0.6
		8	286174	0.174	0.132	224290	0.609	0.569
	2	1	185119	0.407	0.142	182407	0.743	0.698
		2	111081	0.4	0.256	121711	0.86	0.853
		3	128474	0.364	0.22	128610	0.833	0.822
		5	120223	0.424	0.243	129222	0.853	0.84
		6	153947	0.37	0.183	155789	0.78	0.754
		7	179484	0.38	0.151	199578	0.66	0.624
		8	185332	0.381	0.146	201870	0.656	0.615
2	1	1	219248	0.239	0.204	125018	0.801	0.802
		3	186561	0.25	0.227	111772	0.806	0.818
		5	264113	0.21	0.174	129625	0.791	0.783
		6	255218	0.196	0.172	119236	0.774	0.786
		7	247485	0.202	0.17	134938	0.724	0.722
		8	345153	0.196	0.152	165311	0.718	0.683
	2	1	223944	0.413	0.179	143582	0.825	0.791
		2	135570	0.445	0.321	97222	0.913	0.912
		3	155550	0.4	0.275	101534	0.902	0.898
		5	145114	0.458	0.299	104331	0.908	0.904
		6	188431	0.387	0.223	121305	0.87	0.855
		7	229704	0.371	0.18	149358	0.767	0.739
		8	236287	0.372	0.175	150915	0.762	0.729
			Coverage <= Cut off value			Coverage > Cut off value		

Table 5-4. Continued.

Cut off	Run	Lane id	Junctions	Avg Score	Perc Canon	Junctions	Avg Score	Canonical Fraction
5	1	1	259337	0.298	0.268	84929	0.887	0.887
		3	225201	0.318	0.302	73132	0.892	0.898
		5	305148	0.262	0.228	88590	0.881	0.878
		6	296541	0.25	0.233	77913	0.875	0.881
		7	296642	0.251	0.224	85781	0.855	0.852
		8	397801	0.234	0.187	112663	0.83	0.806
	2	1	263656	0.442	0.231	103870	0.907	0.892
		2	169910	0.527	0.427	62882	0.948	0.95
		3	189370	0.475	0.37	67714	0.944	0.945
		5	178189	0.526	0.394	71256	0.947	0.948
		6	223803	0.44	0.295	85933	0.931	0.927
		7	280701	0.397	0.23	98361	0.899	0.885
		8	287818	0.395	0.222	99384	0.898	0.88
10	1	1	289649	0.352	0.326	54617	0.924	0.922
		3	253947	0.377	0.365	44386	0.926	0.926
		5	335371	0.311	0.28	58367	0.921	0.917
		6	326184	0.3	0.286	48270	0.917	0.916
		7	326196	0.296	0.272	56227	0.907	0.902
		8	431187	0.269	0.224	79277	0.887	0.87
	2	1	288984	0.475	0.278	78542	0.938	0.933
		2	197064	0.583	0.497	35728	0.958	0.96
		3	216813	0.532	0.44	40271	0.956	0.957
		5	205938	0.58	0.466	43507	0.958	0.96
		6	251758	0.49	0.36	57978	0.949	0.95
		7	308843	0.434	0.28	70219	0.936	0.93
		8	316415	0.432	0.271	70787	0.935	0.926
			Coverage <= Cut off value			Coverage > Cut off value		

Table 5-5. Average coverage (in reads/junction) and probability score on junctions predicted from Mbnl3 RNA-Seq datasets and their comparisons with ENSEMBL known genes. Only junctions with a maximum coverage of 100 are studied. Known junctions are predicted junctions appearing in the ENSEMBL known genes, and unknown junctions and predicted junctions not appearing in ENSEMBL known genes.

		Known Junctions				Unknown Junctions			
Run	Lane id	Junctions	Avg Cov	Avg Score	Canonical Fraction	Junctions	Avg Cov	Avg Score	Canonical Fraction
1	1	127707	13.393	0.818	0.922	216559	1.872	0.222	0.125
	3	120058	11.559	0.817	0.933	178275	1.857	0.217	0.122
	5	127561	14.353	0.822	0.923	266177	1.808	0.200	0.111
	6	121484	12.362	0.799	0.924	252970	1.737	0.178	0.1
	7	121641	14.724	0.799	0.921	260782	2.033	0.194	0.106
	8	135760	18.913	0.814	0.900	374704	2.151	0.203	0.116
2	1	127102	22.007	0.896	0.933	240424	2.120	0.403	0.146
	2	119385	9.819	0.903	0.961	113407	1.629	0.364	0.155
	3	119715	10.806	0.904	0.960	137369	1.627	0.332	0.139
	5	122562	11.343	0.907	0.960	126883	1.676	0.394	0.158
	6	125888	14.778	0.909	0.955	183848	1.811	0.348	0.138
	7	126222	19.085	0.904	0.946	252840	2.151	0.339	0.127
	8	127674	19.007	0.895	0.932	259528	2.140	0.342	0.124

Table 5-6. Average coverage (in reads/junction) and probability score on junctions by canonical signal using Mbnl3 RNA-Seq datasets. Only junctions with a maximum coverage of 100 are studied. Canonical junctions that predicted junctions with canonical signal and Non-canonical junctions are predicted junctions without the canonical signal.

		Canonical Junctions			Non-canonical Junctions		
Run	Lane id	Junctions	Avg Score	Avg Cov	Junctions	Avg Score	Avg Cov
1	1	144869	0.783	11.85	199397	0.196	2.001
	3	133780	0.787	10.428	164553	0.192	1.967
	5	147430	0.777	12.493	246308	0.177	1.909
	6	137620	0.756	10.955	236834	0.161	1.831
	7	139533	0.756	12.907	242890	0.174	2.142
	8	165421	0.753	15.589	345043	0.179	2.303
2	1	153599	0.871	18.534	213927	0.360	2.151
	2	132287	0.888	8.983	100505	0.315	1.678
	3	133965	0.884	9.781	123119	0.288	1.68
	5	137729	0.89	10.242	111716	0.346	1.721
	6	145698	0.882	12.946	164038	0.304	1.871
	7	151673	0.869	16.193	227389	0.299	2.185
	8	151274	0.864	16.272	235928	0.307	2.207

Table 5-7. Detected alternative 5'/3' splicing sites not appearing in Maize known gene annotations in the three maize RNA-Seq datasets.

	5'/3' Alternative sites by group	Shared alternative sites in two mutants	Shared alternative sites in all samples
FSU RNA-Seq	4324	NA	312
Sample3 RNA-Seq	1900	552	312
Sample4 RNA-Seq	1262	552	312

Table 5-8. Detected exon skips not appearing in maize gene annotations from maize RNA-Seq runs FSU, sample-3 and sample-4.

	5'/3' exon skips by group	Shared exon skips in two mutants	Shared exon skips in all samples
FSU RNA	801	NA	58
Sample3 RNA-Seq	192	83	58
Sample4 RNA-Seq	182	83	58

Table 5-9. Results of de novo gene assembly using PASTA and Cufflinks on two mouse RNA-Seq datasets. Reference genes are taken from ENSEMBL annotations. Sample 1 RNA-Seq dataset contains 49 million paired-end reads of 40 base pairs from the combined RNA-Seq datasets in lane1 of Run1 and Run2. Sample 2 RNA-Seq dataset contains 19 million paired-end reads of 40 base pairs from lane8 of Run2. These two RNA-Seq datasets are generated from two independent experiments of RNA-Seq.

	Num of genes	Num of isoforms	Average length of isoform	Average num of isoforms per Gene	Average num of exons per isoform	Average num of nucleotide per exon
Reference	31804	49039	2540.26	1.54	10.09	253.3
PASTA (sample 1)	30412	37415	968.43	1.23	6.50	146.2
Cufflinks (sample 1)	102667	103651	990.64	1.01	1.43	691.5
PASTA (sample 2)	16064	22868	867.01	1.42	7.73	147.6
Cufflinks (sample 2)	63531	65529	1088.18	1.03	2.06	527.3

Table 5-10. The number of differentially expressed genes reported by MISO, IsoEM and PASTA from Mbnl3 isoform knockout mice. I only take their most strongly regulated genes with a minimum fold change of 2.

	MISO	IsoEM	PASTA	PASTA-MISO	PASTA-IsoEM	MISO-IsoEM	MISO, IsoEM and PASTA
Differentially expressed genes (Run2)	2449	5578	493	59	420	27	6
Differentially expressed genes (Run1)	2783	12207	233	41	224	31	16
Differential expressed genes shared by Run1 and Run2	237	226	65	NA	NA	1	0

Table 5-11. Average exon expression differences between different lanes of Mbnl3 RNA-Seq datasets from Run2.

Lane id	Lane id	Expression difference
1	2	0.29020718
1	3	0.9351454
2	3	0.6513815
5	6	0.21064514
5	7	-0.8648311
5	8	-0.85037327
6	7	-1.0738099
6	8	-1.0577421
7	8	0.015503948

Table 5-12. Differentially expressed isoform in the Mbnl3 RNA-Seq datasets (Run 1 and Run 2) using different software based on a minimum of fold change of 2.

	Transcript ID	Gene name	Fold Change (MISO)	Fold Change (IsoEM)	PASTA
	ENSMUST00000114250	Col11a2	0.003	0.087	
Differentia	ENSMUST00000084319	Med8	0.008	0.063	
lly	ENSMUST00000052999	Echdc2	0.008	0.100	
expressed	ENSMUST00000097995	Mllt3	0.028	0.100	
isoform	ENSMUST00000029676	Adam15	0.030	0.125	
(down	ENSMUST00000095945	Wdhd1	0.069	0.077	
regulation)	ENSMUST00000111309	Mdk	0.077	0.100	
by Run2	ENSMUST00000113120	Bex1	0.115	0.067	
	ENSMUST00000106579	Spata6	0.133	0.143	
	ENSMUST00000105960	Trnau1ap	0.143	0.143	Y
	ENSMUST00000117419	Arhgef1	0.151	0.125	Y
	ENSMUST00000001620	Fxr1	0.159	0.059	
	ENSMUST00000115316	Klhl13	0.167	0.077	
	ENSMUST00000114931	Cuta	0.169	0.091	
	ENSMUST00000112208	Prkcd	0.200	0.091	
Differentia	ENSMUST00000101198	Tmem164	200.000	6.993	
lly	ENSMUST00000113038	Uhrf1	83.333	6.494	Y
expressed	ENSMUST00000120204	Bai2	62.500	6.993	
isoform	ENSMUST00000110463	Cops2	32.258	6.993	
(up	ENSMUST00000109825	H13	16.949	5.988	
regulation)	ENSMUST00000029175	Src	11.905	10.989	
by Run2	ENSMUST00000084996	Odz4	11.765	6.993	
	ENSMUST00000111293	Phf21a	8.000	5.988	Y
	ENSMUST00000034342	Cklf	3.497	8.000	
	ENSMUST00000086047	AW551984	3.145	6.667	
	ENSMUST00000119215	Snrk	3.003	6.494	Y
	ENSMUST00000113005	Cnot6l	2.053	9.009	Y
Differentia	ENSMUST00000113756	Odf2	0.001	0.071	Y
lly	ENSMUST00000116646	Stau2	0.007	0.143	
expressed	ENSMUST00000023053	Col14a1	0.009	0.143	Y
isoform	ENSMUST00000056391	Men1	0.011	0.083	Y
(down	ENSMUST00000062368	Dhx30	0.011	0.125	
regulation)	ENSMUST00000111448	Cugbp1	0.013	0.143	
by Run1	ENSMUST00000092295	Mbd3	0.022	0.143	
	ENSMUST00000060992	Rtn4	0.026	0.143	Y
	ENSMUST00000029582	Col11a1	0.038	0.122	

Table 5-12. Continued.

	Transcript ID	Gene name	Fold Change (MISO)	Fold Change (IsoEM)	PASTA
	ENSMUST00000107213	Nbr1	0.047	0.067	
	ENSMUST00000115804	Pdxdc1	0.062	0.077	Y
	ENSMUST00000054588	Col9a1	0.063	0.046	Y
	ENSMUST00000113377	Golga2	0.149	0.111	Y
	ENSMUST00000105346	Tcfe2a	0.163	0.083	
	ENSMUST00000107346	2500003M10Rik	0.186	0.077	Y
	ENSMUST00000115316	Klhl13	0.200	0.125	Y
	ENSMUST00000112326	Rbbp7	0.250	0.089	Y
	ENSMUST00000112500	Rps6ka3	0.265	0.143	
	ENSMUST00000113673	Drap1	0.289	0.143	
	ENSMUST00000090791	Rprd2	0.290	0.111	
Differentially expressed isoform (up regulation) by Run1	ENSMUST00000050201	Mapk3	500.000	18.868	
	ENSMUST00000115457	Eif4g1	333.333	5.464	Y
	ENSMUST00000109088	Gnas	83.333	8.000	Y
	ENSMUST00000105815	Mul1	25.641	8.333	
	ENSMUST00000094917	Als2cr4	16.949	5.988	Y
	ENSMUST00000071248	Agrn	14.925	14.085	
	ENSMUST00000118987	Bahcc1	8.197	9.009	Y
	ENSMUST00000117450	Bnip2	8.130	5.988	
	ENSMUST00000108797	Dnajc5	8.000	12.987	
	ENSMUST00000115195	Dnajc2	5.128	5.988	Y
	ENSMUST00000120289	Pfn2	4.545	6.993	
	ENSMUST00000063477	Ap1b1	3.584	5.988	Y

Table 5-13. The top 10 differentially expressed genes from Mbnl3 RNA-seq (the combined sample 2010 Runs and 2011 Runs) using different algorithms (Values in IsoEm column and the first two columns of MISO represents the overall fold-changes across all pairwise comparisons between control and mutant, and values in the last MISO column represents Bayes factor—the likelihood of alternative splicing, the last column tells us whether our program predicts this gene or not).

Gene name	IsoEM (Run2)	IsoEM (Run1)	MISO (Run2) Fold Change	MISO (Run1) Fold Change	MISO (Run2)	MISO (Run1)	Regulation by IsoEM	PASTA
Terf2	1.2e-2	9e-3	2.2e-1	1	1e+12	3.9e+69	Up	
Ablim1	263.40	225.83	1	1	2.7e+159	6.3e+18	Down	Y
Tnnt3	201.21	23303.19	9.5e-1	0.06	4.6e+53	6.7e+60	Down	Y
Vegfa	144.70	7.716	2.7e-6	1 *	6.3e+70	9.1e+52	Down	Y
Foxp4	152.59	1201.77	2.6	0.35	7e+30	1.8e+241	Down	
Mapk14	6.4e-4	1.2e-3	1	1	4.5e+116	2.2e+257	Up	Y
Epb4.1	1.3e-4	3.1e-3	1	1	1.4e+306	2.2e+248	Up	Y
Tpm1	2.6e-3	7.6e-6	1	1	9e+94	3.3e+34	Up	Y
Klc1	4.9e-4	4.3e-4	1.3e-1	0.25	1.6e+150	7e+70	Up	Y
Hmga1	4244.0	25277.38	1	1	4.3e+204	1e+13	Down	Y
Rhgef1	9.9e-2	4.5e-5	1	1	3.e+66	1e+65	Down	Y

Table 5-14. Detection of differentially expressed isoforms using the Genetic algorithm on mouse Mbnl3 RNA-Seq datasets from Run2.

Gene	mRNA accession	Control sample ID		Mutant sample ID				Control Avg	Mutant Avg	Fold-change	Notes
		1	3	5	6	7	8				
Itpr2	NM_019923	0.65	0.83	2.42	2.08	2.31	3.22	0.74	2.51	3.41	Exon skip
	NM_010586	1.50	1.45	0.52	0.87	0.31	1.53	1.47	0.81	0.55	
Zc3h7a	NR_027502	5.89	5.81	2.09	3.89	1.24	1.12	5.85	2.08	0.36	Mutually exclusive exons
	NM_145931	10.0	8.97	16.07	12.11	12.99	29.88	9.49	17.76	1.87	
Cnnm2	NM_001102471	3.13	2.24	1.61	0.65	3.37	3.11	2.69	2.19	0.81	Exon skip
	NM_033569	0.43	0.38	3.03	1.79	0.69	2.30	0.41	1.95	4.80	

Table 5-15. Number of reads and junctions detected using Mbnl3 RNA-Seq datasets. This table displays the total number of reads, the total number of junctions identified by both programs, and the ratio between these two numbers for Run 1 and Run2 respectively

	Lane id	Number of reads (millions)	Number of aligned reads (millions)	PASTA junction clusters	PASTA junctions selected
Control	1	18.15	7.68	932236	160285
	2	23.90	11.23	924742	178095
	3	27.84	12.90	1142055	196082
Mutant	5	21.20	9.60	949623	174676
	6	21.71	9.79	959808	174821
	7	8.30	4.15	284701	93353
	8	19.01	9.97	317695	164215

Table 5-16. Average coverage (in reads/junction) and probability score on junctions by canonical signal using Mbnl2 RNA-Seq datasets. Only junctions with a maximum coverage of 100 are studied. Canonical junctions that predicted junctions with canonical signal and Non-canonical junctions are predicted junctions without the canonical signal.

	Canonical Junctions			Non-canonical Junctions		
Lane id	Junctions	Avg Score	Coverage	Junctions	Avg Score	Coverage
1	151735	0.748	12.250	780501	0.313	1.543
2	161160	0.770	15.983	763582	0.322	1.730
3	177184	0.752	17.028	964871	0.313	1.664
5	158903	0.764	13.750	790720	0.315	1.647
6	159808	0.760	14.043	800000	0.315	1.677
7	85779	0.809	9.633	198922	0.340	1.982
8	143771	0.877	15.941	173924	0.456	2.632

Table 5-17. The number of differentially expressed isoforms reported by MISO, IsoEM and PASTA using Mbnl2 RNA-Seq datasets.

	MISO >1.3 fold	MISO >1.5 fold	IsoEM >1.3 fold	IsoEM >1.5 fold	MISO - IsoEM	PASTA >1.5 fold	MISO, IsoEM and PASTA
Differentially expressed isoforms (Up-regulation)	619	425	5772	3387	69	2813 (both up and down regulation)	81 (both up and down regulation)
Differentially expressed isoforms (Down-regulation)	1453	1030	4781	2961	168		

Table 5-18. Gene ontology analysis of genes with differentially expressed isoforms from MISO, IsoEM and PASTA analysis using Mbnl2 RNA-Seq datasets.

Gene Ontology	Candidate genes
tRNA-intron endonuclease activity	Tsen34
endoplasmic reticulum	Ube2j2
catalytic activity	Steap2
actin binding	Ablim1, Synpo2
nucleic acid binding	Zfp536
Lysosome	March1
metalloendopeptidase activity	Adam22
plasma membrane	Slc43a2
postsynaptic membrane organization	Arhgef9
calcium ion binding	S100a16
protein binding	Pcdhgb4, Shank2, Snap91, Sh3pxd2a
Binding	Gprasp1
structural molecule activity	Col15a1" "Col12a1
stress fiber	Sorbs1
Cytoplasm	Atp6v0a1
guanyl-nucleotide exchange factor activity	Rasgrp2
cellular_component	1810014F10Rik, Aarsd1
histone deacetylase complex	Arhgef10l
molecular_function	Klhl13, Ralgps1, Zfp385b, Dcun1d2, Osbpl1a, Gpr155, Btf3l4, C530008M17Rik Ttc39b, B230312A22Rik, Dcun1d3 Chchd7, Zfp740, D19Wsu162e, Macrod2 Zfp668, 0610037L13Rik, Rap1gap 1110051M20Rik, Trim37, Mical2, Fam53b Clec16a, Tmem164, Yeats2, 2310001H12Rik, Echdc2, Scaper 4932438A13Rik, Phactr2, Lhfpl2
skeletal muscle contraction	Homer1
blood vessel development	Rapgef1
ubiquitin thiolesterase activity	Otud7b
Golgi membrane	Arfgap1
GTPase activator activity	Tbc1d1
creatine transporter activity	Slc6a8
receptor activity	Sema6d, Slc22a17, Sema3b, Sema6a
DNA binding	Tcf7l2
Rho guanyl-nucleotide exchange factor activity	Plekhg5
transmembrane receptor protein tyrosine kinase	Sh2d3c
adaptor protein activity	Camk1d, Pip5k1a, Hnrnph2, Elavl4 Myo18a,
nucleotide binding	Pde10a

Table 5-19. Alternative junctions from Mbnl2 RNA-Seq datasets identified by PASTA.

gene	chromosome	start	end	Fold (>=1.5)	T-Test (p=0.05)	KS-Test (p=0.064)	GA	Splicing event
Tanc2	chr11	105774803	105776205	Y		Y	Y	alt_cassette
Kcnma1	chr14	24150192	24156166	Y	Y	Y	Y	alt_cassette
Gas5	chr1	162967295	162967418	Y				retained_int
Limch1	chr5	67365901	67388358	Y		Y		alt_cassette
Clasp2	chr9	113771405	113777665	Y		Y	Y	twin_cassett
Cxadr	chr16	78329403	78336616		Y	Y		alt_end
Cxadr	chr16	78336616	78360030		Y	Y		alt_end
Spna2	chr2	29855562	29857771	Y		Y		alt_cassette
Dgkh	chr14	78969964	78977255	Y	Y	Y		alt_cassette
Arpp21	chr9	112088273	112090093			Y	Y	alt_cassette
Mapre3	chr5	31165713	31166967	Y	Y	Y		alt_3
St3gal3	chr4	117704487	117780153	Y		Y		alt_cassette
Cyfip1	chr7	63153342	63155284	Y		Y		alt_cassette
Ndrg4	chr8	98235044	98237077	Y				alt_cassette
Cyfip1	chr7	63152185	63153607	Y		Y		alt_cassette
Csnk1d	chr11	120825220	120829308	Y	Y	Y	Y	alt_cassette
Ppp1r12a	chr10	107689980	107696811	Y		Y		alt_cassette
Cacna1d	chr14	30941984	30943302	Y	Y	Y	Y	alt_cassette
Mbnl1	chr3	60417442	60419541			Y		alt_cassette
Add1	chr5	34971162	34973118	Y				alt_cassette
Hivep2	chr10	13786490	13837324	Y				alt_cassette
Dst	chr1	34356683	34359674	Y		Y	Y	alt_cassette
Tom1l2	chr11	60044639	60055319	Y		Y		alt_cassette
Ablim2	chr5	36179746	36183929	Y	Y	Y		alt_cassette
Bbx	chr16	50209257	50220557			Y		alt_cassette
Slain2	chr5	73349525	73365769	Y				alt_cassette
Spint2	chr7	30044471	30045389	Y			Y	alt_cassette
Mbnl2	chr14	120788385	120795672	Y	Y	Y	Y	alt_cassette
Ank3	chr10	69413806	69418673	Y	Y	Y		alt_cassette
Sgip1	chr4	102591460	102600065	Y	Y	Y	Y	alt_cassette
Dlg2	chr7	99535500	99576200	Y		Y	Y	complex
Dlgap1	chr17	71012149	71110414	Y		Y		complex
Jag2	chr12	114153708	114154111	Y	Y	Y		alt_cassette
Ppm2c	chr4	11889380	11893448	Y				alt_cassette
Inpp4a	chr1	37455745	37465342	Y	Y	Y		complex
Hisppd2a	chr2	121147528	121153380	Y	Y	Y	Y	alt_cassette
Nmt2	chr2	3222219	3222668	Y	Y		Y	retained_int
Arhgap26	chr18	39152798	39535939	Y		Y	Y	Ncomplex
Atp8a1	chr5	68170272	68177864	Y	Y	Y	Y	complex
Cacna1d	chr14	30896182	30908503	Y		Y	Y	complex
Aktip	chr8	93655027	93657819	Y		Y		complex

Table 5-19. Continued

gene	chromosome	start	end	Fold (>=1.5)	T-Test (p=0.05)	KS-Test (p=0.064)	GA	Splicing event
Dnm1	chr2	32164584	32167646	Y		Y	Y	complex
Bmp1	chr14	70886079	70889826	Y				complex
Ptprg	chr14	12999371	13023216	Y	Y	Y		alt_cassette
Agap2	chr10	126526414	126527482					alt_cassette
Ccnd2	chr6	127100818	127100884					retained_int
Sh3kbp1	chrX	156254638	156266305	Y		Y		twin_cassett
Map4k3	chr17	81023454	81029755	Y		Y		alt_cassette
Pfkp	chr13	6587899	6597984					mutually_exc
Ilf3	chr9	21204457	21209805	Y		Y		alt_end
Anxa6	chr11	54792612	54804892	Y		Y		alt_end
Ep400	chr5	111168700	111170988	Y				alt_cassette
Kif21a	chr15	90774332	90783141	Y		Y	Y	complex
Clta	chr4	44043140	44045272	Y				alt_cassette
Fam126b	chr1	58587103	58592233	Y				alt_cassette
Mapt	chr11	104143794	104156172	Y	Y	Y	Y	twin_cassett
Ganab	chr19	8981884	8983167	Y	Y	Y		alt_cassette
Exoc1	chr5	76983230	76988028	Y	Y	Y		alt_cassette
Arl2bp	chr8	97190654	97191496	Y				alt_start
Bnip3l	chr14	67606612	67607971	Y				alt_cassette
Nrcam	chr12	45645882	45648218	Y		Y	Y	alt_cassette
Grin1	chr2	25165995	25168947	Y				alt_cassette
Kif1a	chr1	94922312	94933422	Y		Y	Y	complex
Insl3	chr8	74204696	74204781					retained_int
Prpf18	chr2	4566699	4569307					alt_cassette
4930506M07Rik	chr19	59049698	59074382					twin_cassett
Lass5	chr15	99570141	99576340	Y				twin_cassett
Nmt2	chr2	3222711	3226738	Y				alt_cassette
Rabgap1l	chr1	162149304	162187573	Y	Y	Y	Y	alt_end
Sept11	chr5	93596713	93602498			Y		alt_3
Clip1	chr5	124081178	124090502	Y				alt_cassette
Arhgef7	chr8	11822029	11831513					complex
Cask	chrX	13128108	13132003	Y		Y		complex
Ica1	chr6	8603535	8606361	Y				alt_cassette
Slc4a7	chr14	15607942	15611339	Y				alt_end
Synj2	chr17	6019478	6044290	Y	Y	Y	Y	complex
Sorbs1	chr19	40448107	40457010	Y	Y	Y		alt_cassette
Mta1	chr12	114355292	114359037	Y				alt_cassette
Myh10	chr11	68558921	68581874	Y				alt_end
Adam23	chr1	63619572	63639181	Y				alt_cassette

Table 5-19. Continued

gene	chromosome	start	end	Fold (>= 1.5)	T-Test (p=0.05)	KS-Test (p=0.064)	GA	Splicing event
A2bp1	chr16	7353157	7375997	Y		Y		mutually_exc
Mapk9	chr11	49686381	49691753				Y	complex
Ap1gbp1	chr11	83852803	83854319	Y	Y	Y	Y	alt_cassette
Tbp	chr17	15636988	15639928	Y				alt_cassette
Spr	chr6	85086925	85087432	Y				alt_3
Plekhg5	chr4	151482439	151482511	Y		Y	Y	retained_int
Sept11	chr5	93602563	93604429	Y		Y		alt_end
Pdlim7	chr13	55608484	55608690	Y	Y	Y	Y	retained_int
Arhgef11	chr3	87521442	87526278		Y	Y		alt_cassette
Camkk2	chr5	123184254	123187441	Y		Y		alt_cassette
Ptprd	chr4	75971229	77857693	Y	Y	Y	Y	alt_start
Ghitm	chr14	37946989	37948452					alt_start
Hp1bp3	chr4	137772655	137778003	Y	Y	Y	Y	alt_cassette
uc007zav-1	chr16	34220333	34227112	Y				alt_3
Depdc5	chr5	33326679	33333390	Y				alt_cassette
Tacc2	chr7	137878664	137883325	Y	Y	Y		alt_5
Sorbs2	chr8	46821556	46830773	Y		Y	Y	alt_cassette
Dclk2	chr3	86595916	86602932	Y	Y	Y	Y	alt_end
Dgkg	chr16	22571836	22575384	Y				alt_cassette
Zfp275	chrX	70590251	70597147	Y				alt_cassette

Table 5-20. Targets for validation using PASTA splice junction predictions from RNA-Seq data.

Gene	Chromosome	Exon start	Exon end	Type
Grin1	chr2	25165995	25168948	skip
Rnf114	chr2	167329101	167336376	skip
Hp1bp3	chr4	137796652	137797361	GC-AG
Clta	chr4	44038432	44045273	skip
Mapt	chr11	104143794	104156173	skip
Dnm1	chr2	32183567	32191296	Mutually exclusive exons
Csnk1d	chr11	120825220	120829309	skip
Ndrg4	chr8	98235044	98237078	skip
Spint2	chr7	30045449	30066531	skip
Ncdn	chr4	126429325	126430569	skip
Vti1b	chr12	80266126	80273207	skip
Dync1li2	chr8	106947130	106949339	skip
Fubp1	chr3	151884372	151885048	GC-AG
Sept8	chr11	53348130	53348451	GC-AG
Pkp4	chr2	59183050	59185777	GC-AG
Gars	chr6	55018162	55019335	AT-AC
Nat15	chr16	3900877	3901783	AT-AC
Hnrnpl	chr7	29599306	29600118	AT-AC
Setd1a	chr7	134940212	134940588	AT-AC
Hsbp1	chr8	121868902	121869513	AT-AC
4930506M07Rik	chr19	59049698	59074383	skip
Slain2	chr5	73349525	73365770	skip
Pdlim7	chr13	55608985	55610141	skip

Table 5-21. List of primers for PCR validation using PASTA predicted splice junctions.

Target Name	Forward	Reverse	Tm
Grin1	ATCAGCAGAGCCGTCACATT	GTCCAGCGTCTGGTTTGAGA	60°C
Rnf114	CAGGTGCCCTGCGGAC	GGTGGCCTTTACACCTTCCA	60°C
Hp1bp3	AAGAATCCGGTGGCTCTGAC	TTGGGACTTGGCTGGTGTTT	60°C
Clta	CCTCGATGCCAATTCTCGGA	CTTTGGCCTGTTTGCTGGAC	60°C
Mapt	TCGCCAGGAGTTTGACACAA	GTCTCCGATGCCTGCTTCTT	60°C
Dnm1	AGACGCGGGTATTGCTGC	CTTCCATGAACGCTTCCCCT	60°C
Csnk1d	GCACGACAGACTGAAGACCA	CCTCATCTGATCTCACGGGC	60°C
Ndrg4	CCAAACTGGACCCAACCACT	CATCCCCTCACTGCTGTCTG	60°C
Spint2	CTGGCCATGTCATCAGTGGT	GTAGCTTCGCTGCTCCTCTC	60°C
Ncdn	ATGATGCCCGCCTTGCC	CAGCCTGGAGCCAAGAGTG	60°C
Vti1b	TCTTCCTCCATCTCTGCCAA	CCTCCTCCGAGCATTTCGAG	60°C
Dync1li2	AAGCTGGCCAATACCCCTTC	GGAGGATGAGCAGGTGTTCC	60°C
Fubp1	GTCGAGGACGAGGTAGAGGT	GTGGAGTGCCCCGAATTGTA	60°C
Sept8	CTGACCATCGTGGATGCTGT	CAAACTGCGCGTCGATGTAG	60°C
Pkp4	TCCTGTCCGATGAAACCGTG	GGTGGACAGAGAAGGGTGTG	60°C
Gars	TGTTGGATGTGCTGACCGTT	TGTAGCACTCATCACAGGCG	60°C
Nat15	CGAGGGGTCCTCAAAGATGG	CCCAGGTGCTGGATGTAGTC	60°C
Hnrnpl	GAGCGTAAACAGCGTGCTTC	GGGTCACCTTGTCCACTGAG	60°C
Setd1a	CCTCCTTCCTTTGAGCCGAG	TCTTTTGCGCTTTGGAGTGC	60°C
Hsbp1	AGACCATGCAGGACATCACC	GTCGTCAATCCGACTGCTCA	60°C
4930506M07Rik	TTGGGGTTCGTGTGTGGAAA	GGCAAAGCCAGACTCTCTCA	60°C
Slain2	ATCCCGAACATCGAGCACAC	AAACGCTTTCACTGGTTGCC	60°C
Pdlim7	GCCAGTCTTCGGTATCCTCC	AGCCTGTTCAGAGCAAACCA	60°C

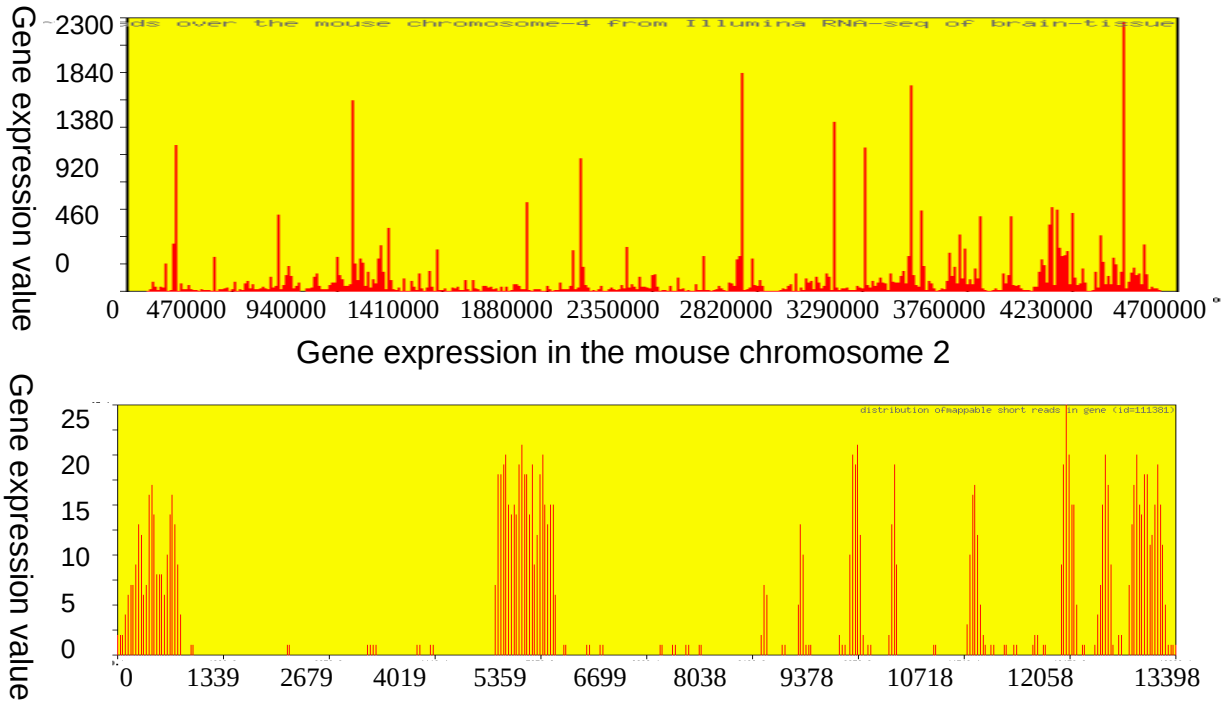


Figure 5-1. RNA-Seq read mapping in chromosome 2 of mouse reference genome build 'mm9'. The upper panel shows the read mapping information from the whole chromosome. The lower panel shows a small chromosome region where a possible transcript is present (as compared with low background noise). The RNA-Seq dataset is downloaded from NCBI SRA via accession number SRX00351.

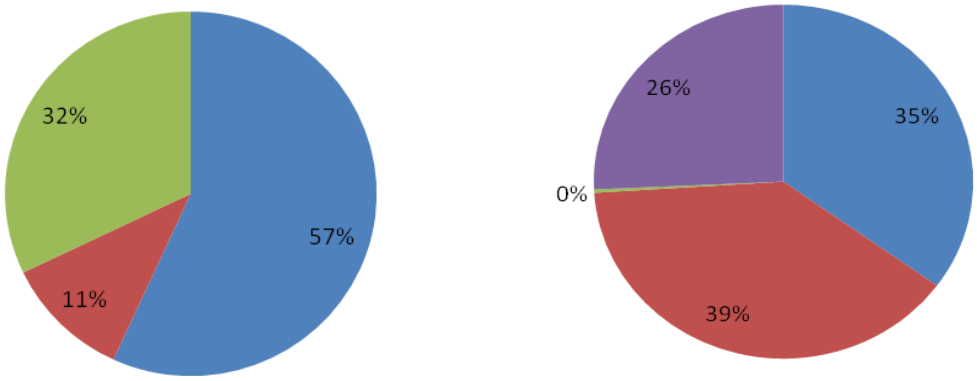


Figure 5-2. Overview of read mapping information using RNA-Seq datasets. The RNA-Seq dataset is downloaded from NCBI SRA via accession number SRX00351. Left panel shows that 57% of reads can be mapped into reference genome, and 11% of them map to splice junctions. The data for right panel comes from the splice junctions. Over all splice junctions. 35% of splice junctions is outside annotated transcript region. 39% of splice junctions are known annotated exon junctions, and 0.03% of them are new alternative splice junctions.

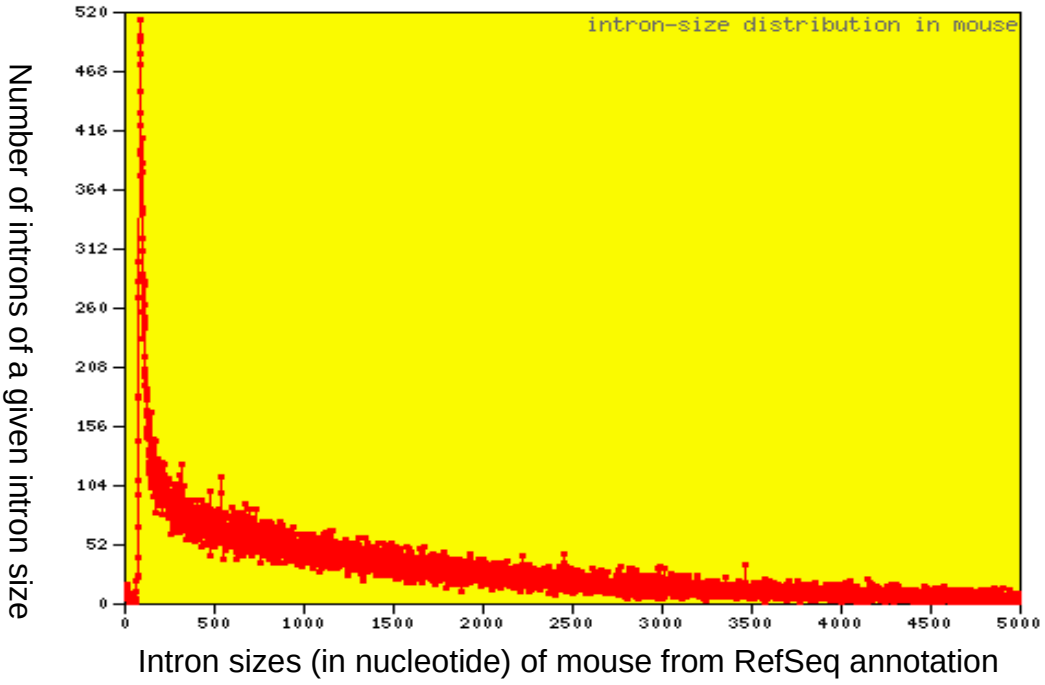


Figure 5-3. Distribution of observed occurrences by different intron sizes from mouse RefSeq gene annotation (introns greater than 5000 nucleotides are not shown). From left to right, there is an increase from around size of 60 and reach the peak with an intron size of 83 nucleotides, followed by a steady decrease with increased intron-size.

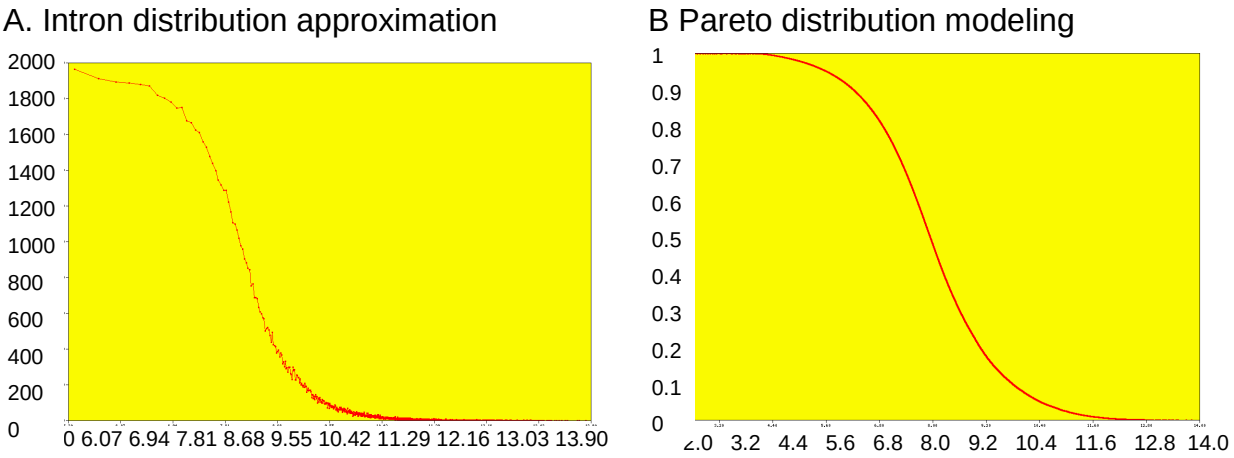


Figure 5-4. A) Distribution of number of occurrences for each mouse intron sizes using UCSC gene annotations. It shows the number of occurrences for each log-normalized intron size. B). Logarithm transformation of intron sizes from mouse RefSeq annotation. The second panels show the number of occurrences within each range of 100 base pairs by normalized intron size. The third panel is the same as the second except the first outline is removed. The fourth panel shows the pareto-value distribution given the logarithm normalized intron sizes.

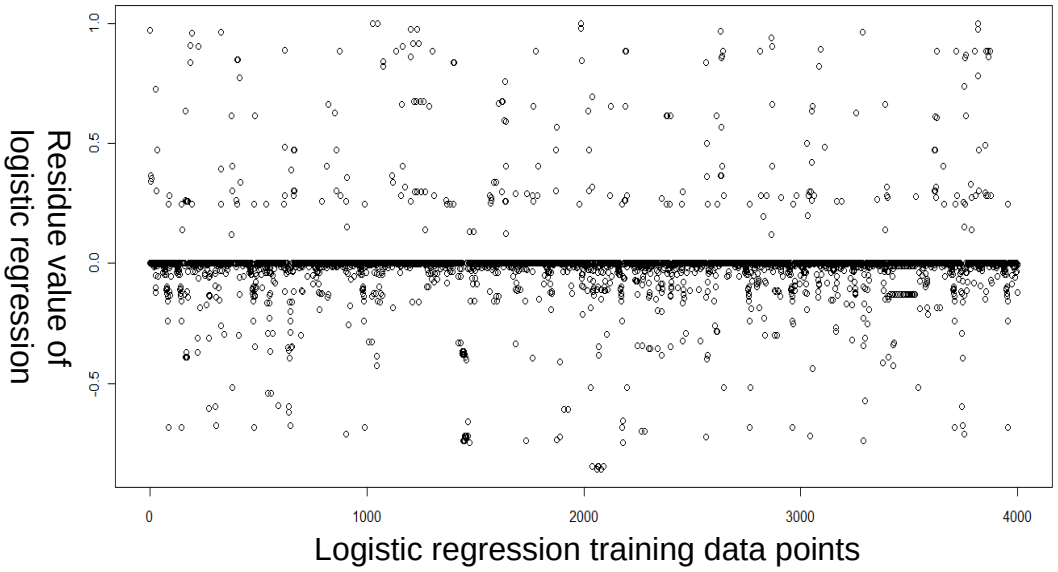


Figure 5-5. The residual distribution of PASTA's logistic regression model using maize RNA-Seq datasets. A total of 4000 points are used for the residual analysis.

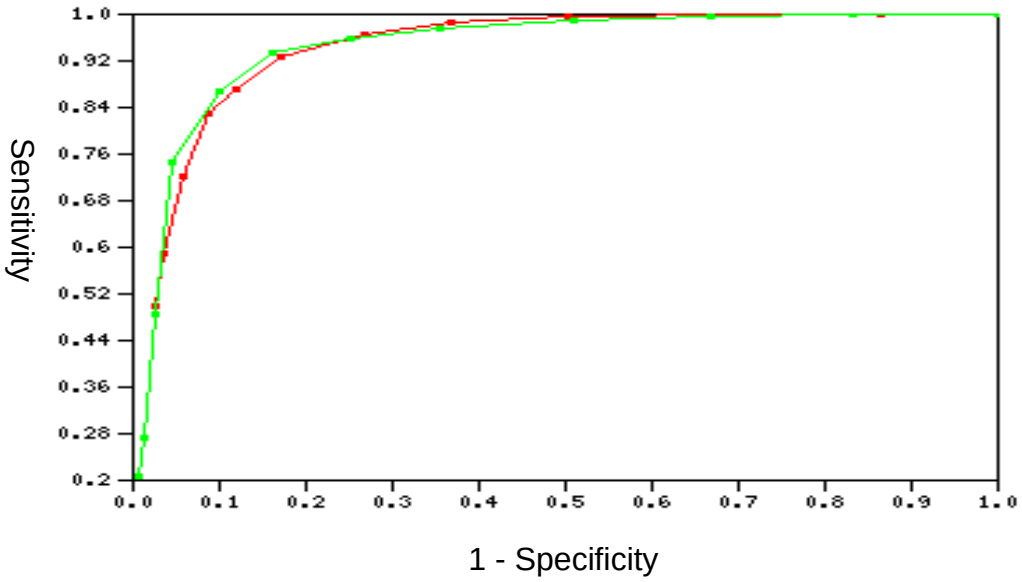


Figure 5-6. PASTA prediction sensitivity and specificity curve by scoring scheme. Green line represent method with consideration of “GT-AG” Splice site, and red line represents method without spice signal.

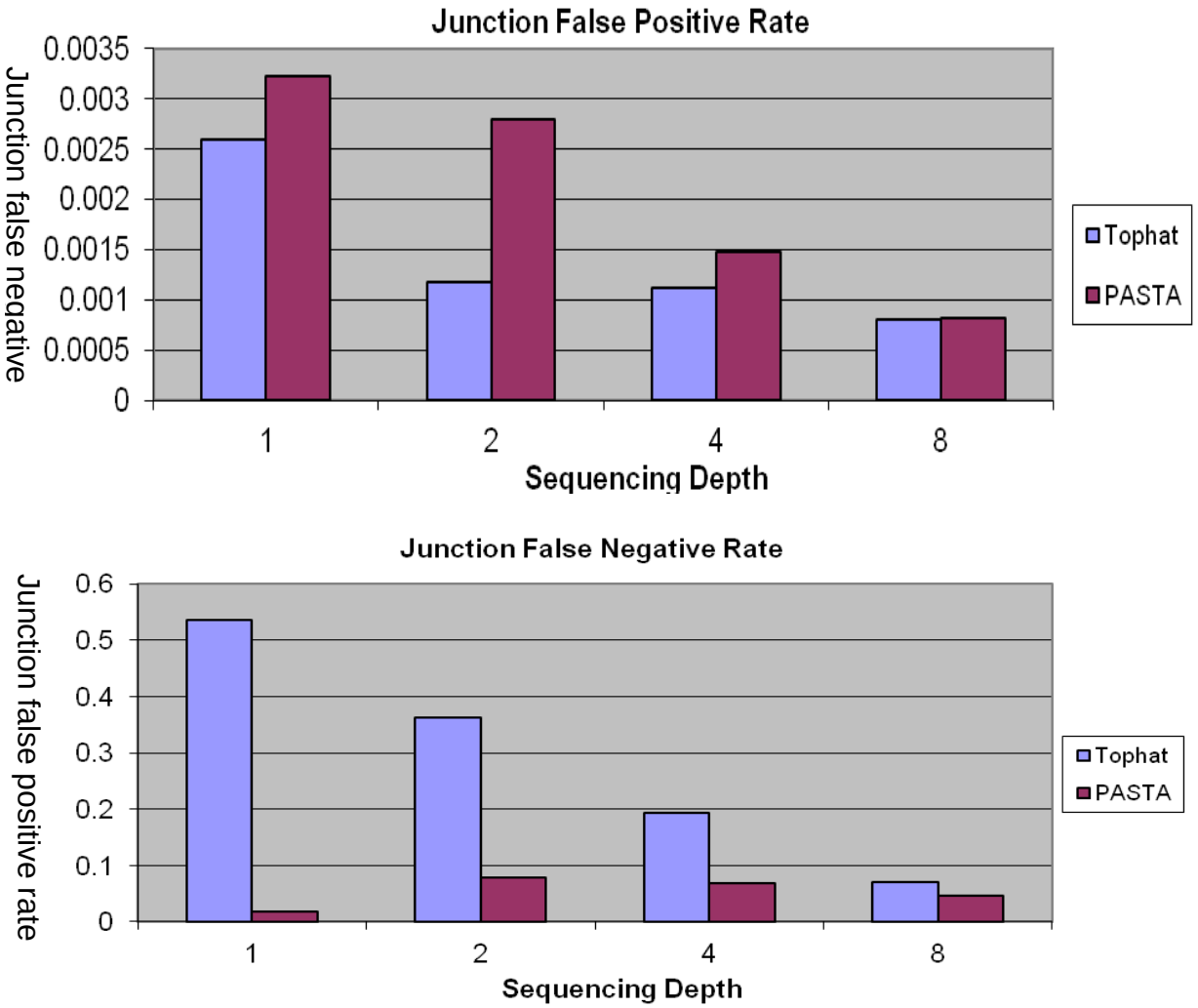


Figure 5-7. Junction accuracy of TopHat and PASTA. Blue bar represents TopHat predictions and red bar represents PASTA predictions. Junction FN rates are shown in the upper panel and junction FP rates are shown in the lower panel. A total of 4 different sequencing depths are simulated.

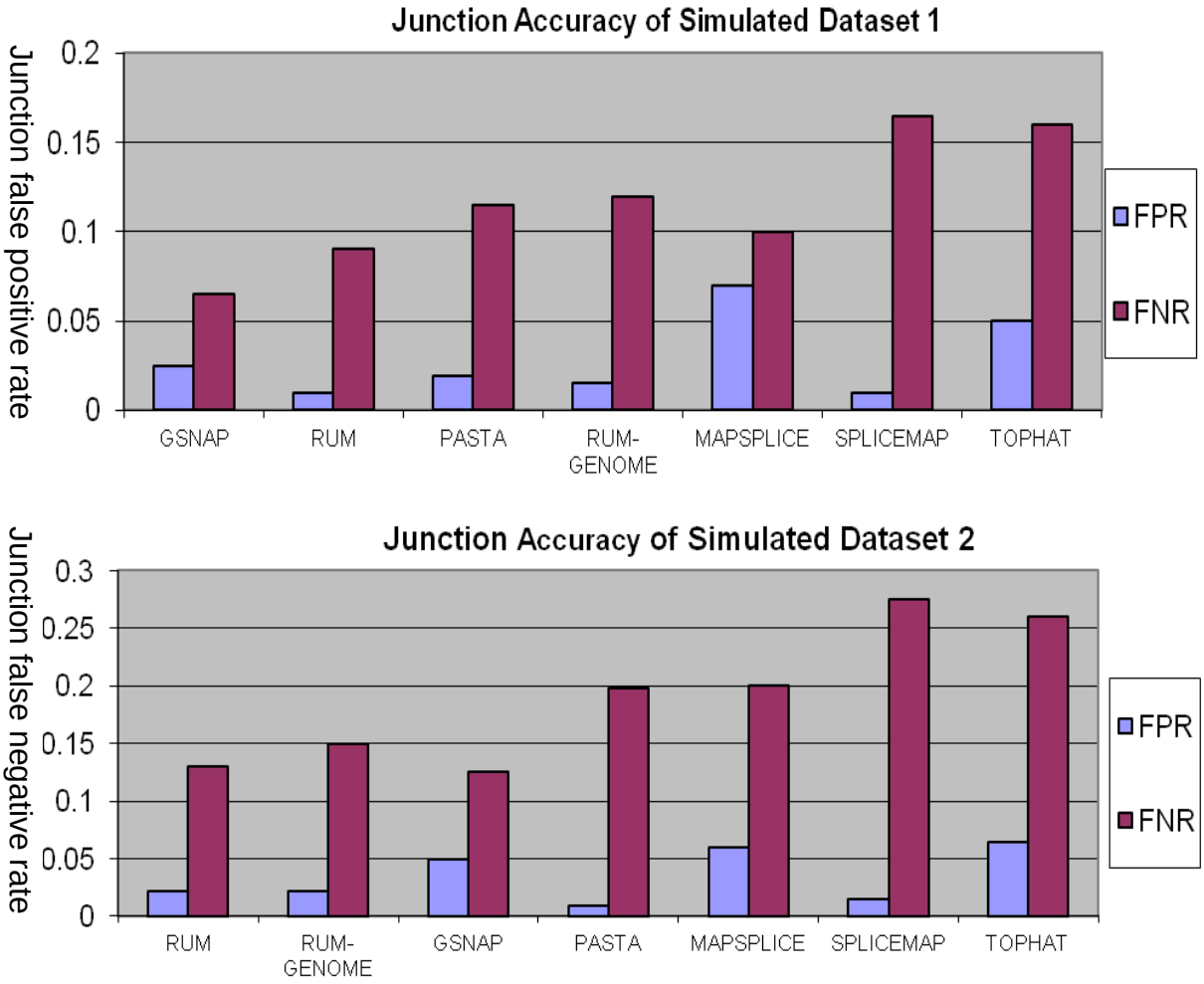


Figure 5-8. Junction accuracy of PASTA and other software. Two simulations are performed, which are shown in the two panels respectively. In the first simulation, the modeling percentage for indel, substitution and error frequency is 0.05%, 0.1% and 0.5% respectively. 80% of the splice signals are coming from annotated splice forms. In the second simulation, the modeling percentage for indel, substitution and error frequency is 0.25%, 0.5% and 0.1% respectively. 65% of the splice signals are coming from annotated splice forms and there are 50% errors in 25% of the trailing 10 bases. The simulation datasets are kindly provided by the Grant lab.

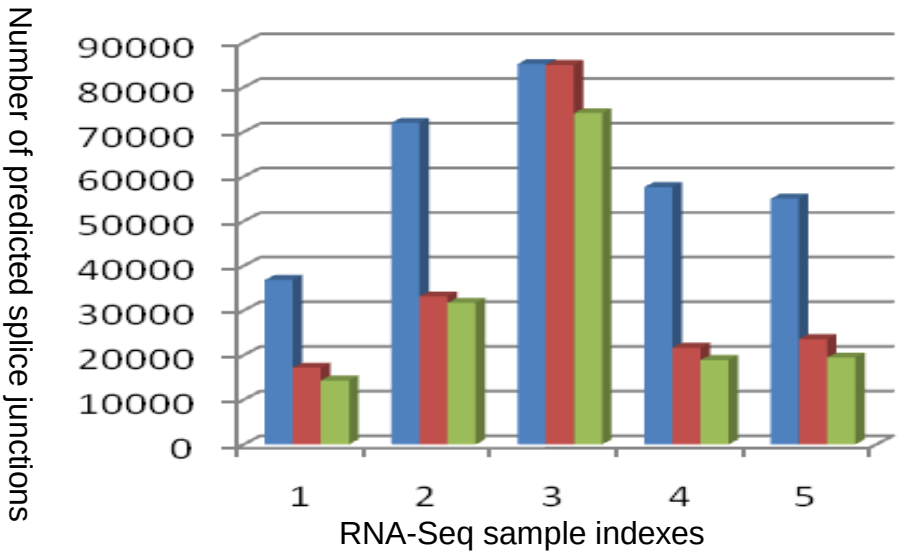


Figure 5-9. Comparison of PASTA and Tophat on splice junction identification using different RNA-Seq samples. The blue color bars contain PASTA predictions, the red color bars show Tophat predictions and the green color bars show the number of splice junction observed both in PASTA and Tophat. The sample indexes are shown in X axis. Index 1 showed results from Illumina paired-end RNA-Seq data (SRA accession number: SRR027863); Index 2 showed results from human Illumina paired-end RNA-Seq data (SRA accession number: SRR027884); Index 3 is from FSU-sample maize Illumina single-end RNA-Seq datasets; Index 4 and 5 are from Sample-3 and Sample-4 of the maize single-end RNA-Seq datasets.

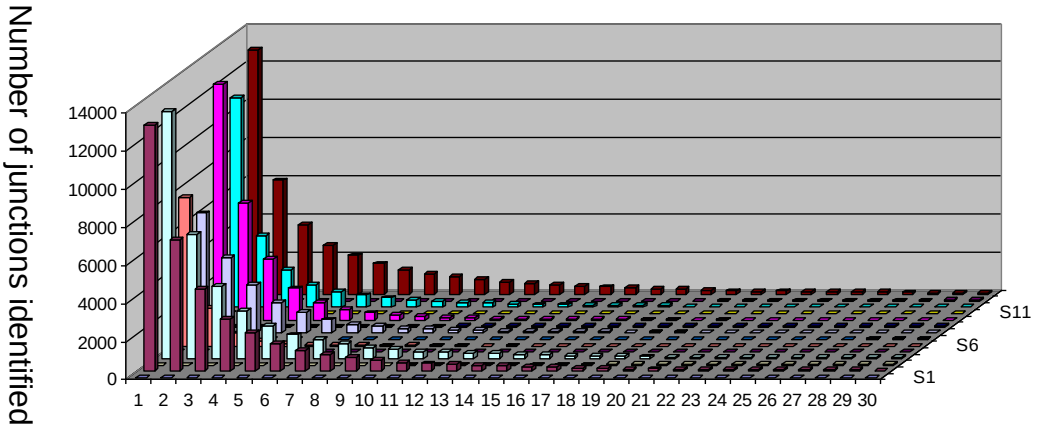


Figure 5-10. Distribution of RefSeq annotated splice junctions that are uniquely predicted by PASTA. Results are supported by Mbnl3 RNA-Seq Run1 and Run2. Each bar represents the total number of splice events that are supported by x reads as indicated in x-axis. A total of 14 experiments (lanes) are shown by the horizontal axis z.

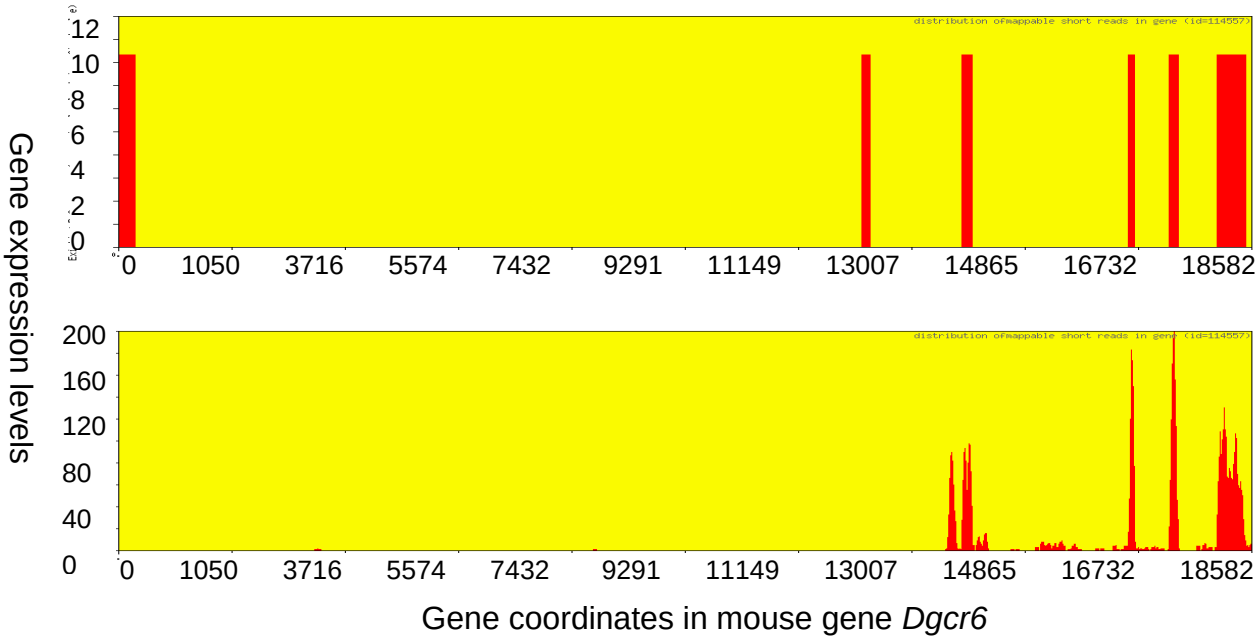


Figure 5-11. Gene expression with RefSeq annotation. The exon coordinates from RefSeq annotation is shown in parallel in the upper panel (with simulated expression level of 10). Gene expression estimates from Mbnl3 RNA-Seq datasets Run2 are displayed in the lower panel. These two figures contain identical chromosome location.

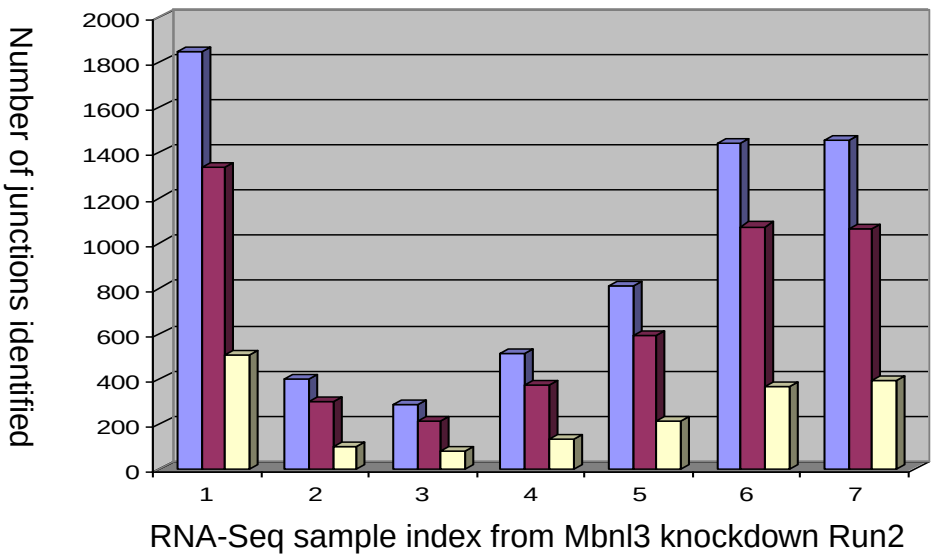


Figure 5-12. Analysis of alternative spliced junctions using Mnbl3 isoform knockout experiment Run2. Control sample lanes are labeled from 1 to 3 and mutant sample lanes are labeled from 4 to 7. The blue bars showed total alternative splice events. Red bars represent alternative sites and yellow bars represent exon skips.

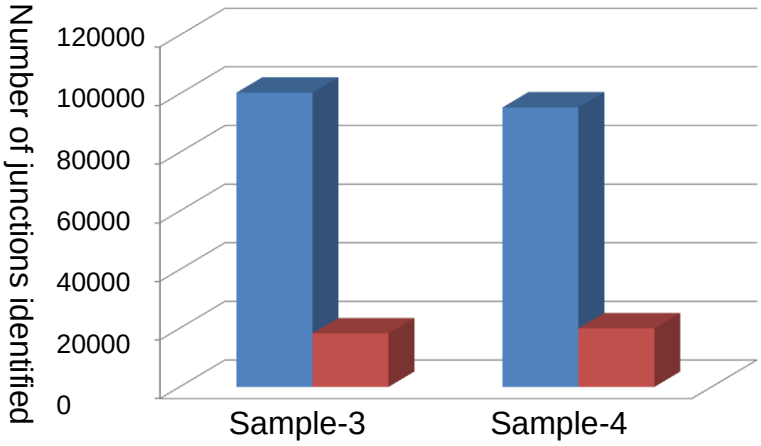


Figure 5-13. Comparisons of PASTA and Tophat in reporting novel splice junctions using Maize RNA-Seq datasets. Blue bars represent splice junctions reported by PASTA, and red bars show putative novel junction identified by Tophat.

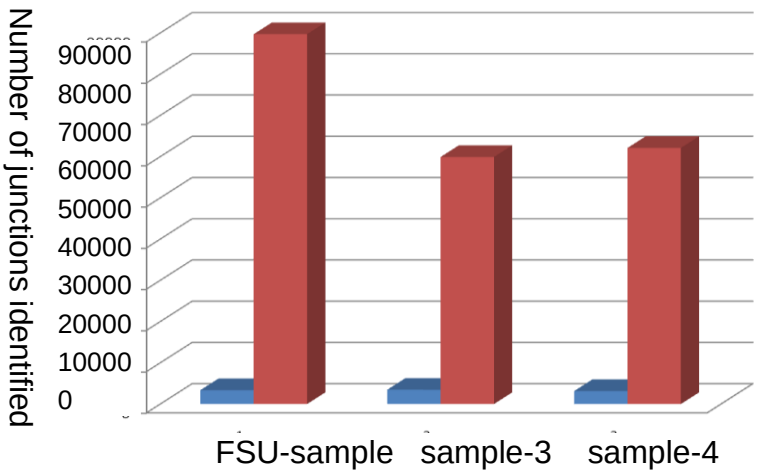


Figure 5-14. The number of splice signals from predicted junctions using three maize RNA-Seq datasets (each junction is supported by at least three reads). Index 1 to 3 is for maize RNA-Seq of FSU, sample-3 and sample-4 respectively. Red bars represent total number of splice junctions identified. Blue bars represent the total number of non-canonical splice junction identified. In total, approximately 3.7%, 5.7% and 5.0% of identified splice clusters are non-canonical splice signals respectively from the three maize RNA-Seq runs.

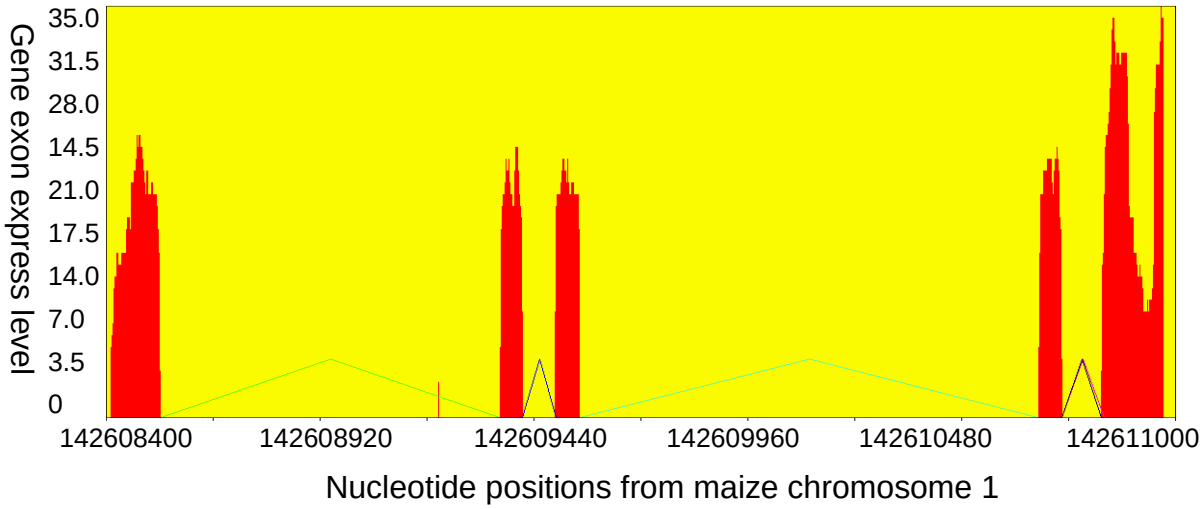


Figure 5-15. A putative novel maize gene using RNA-Seq datasets from Maize RNA-Seq FSU-sample in chromosome 1 from Maize reference sequence assembly.

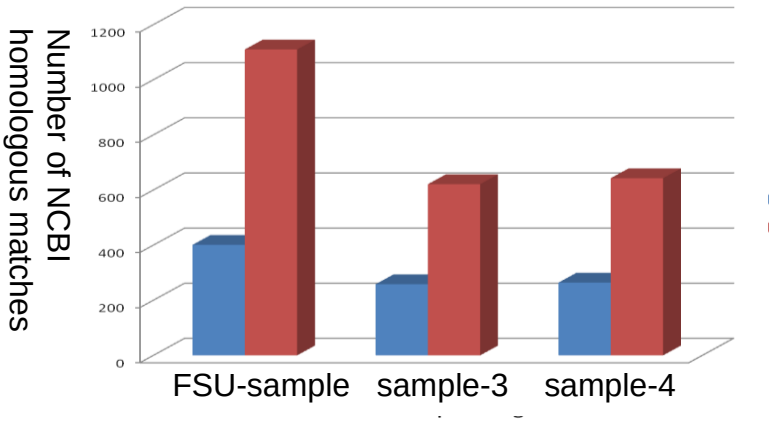


Figure 5-16. Predictions of novel transcripts and their homologous peptide sequences from maize RNA-Seq datasets using NCBI Non-redundant protein database. The red color bar represents putative transcripts not appearing Maize annotation. The blue bar represents the number of peptides from the putative transcripts that are found in NCBI protein database.

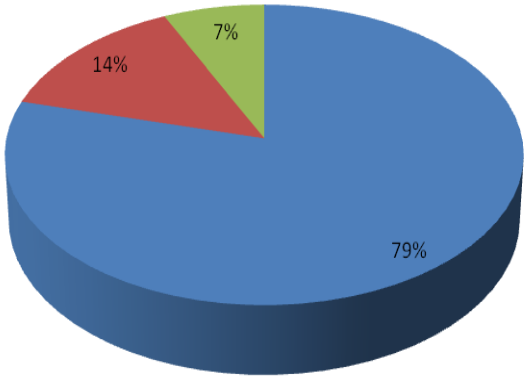


Figure 5-17. The categories of gene isoform assemblies using Mnbl3 RNA-Seq dataset Run2 as compared with RefSeq mouse gene annotations. The blue color portion represents RefSeq annotated genes. Red color portion represents annotated genes that have alternative spliced sites. Green color portion represents putative novel transcripts.

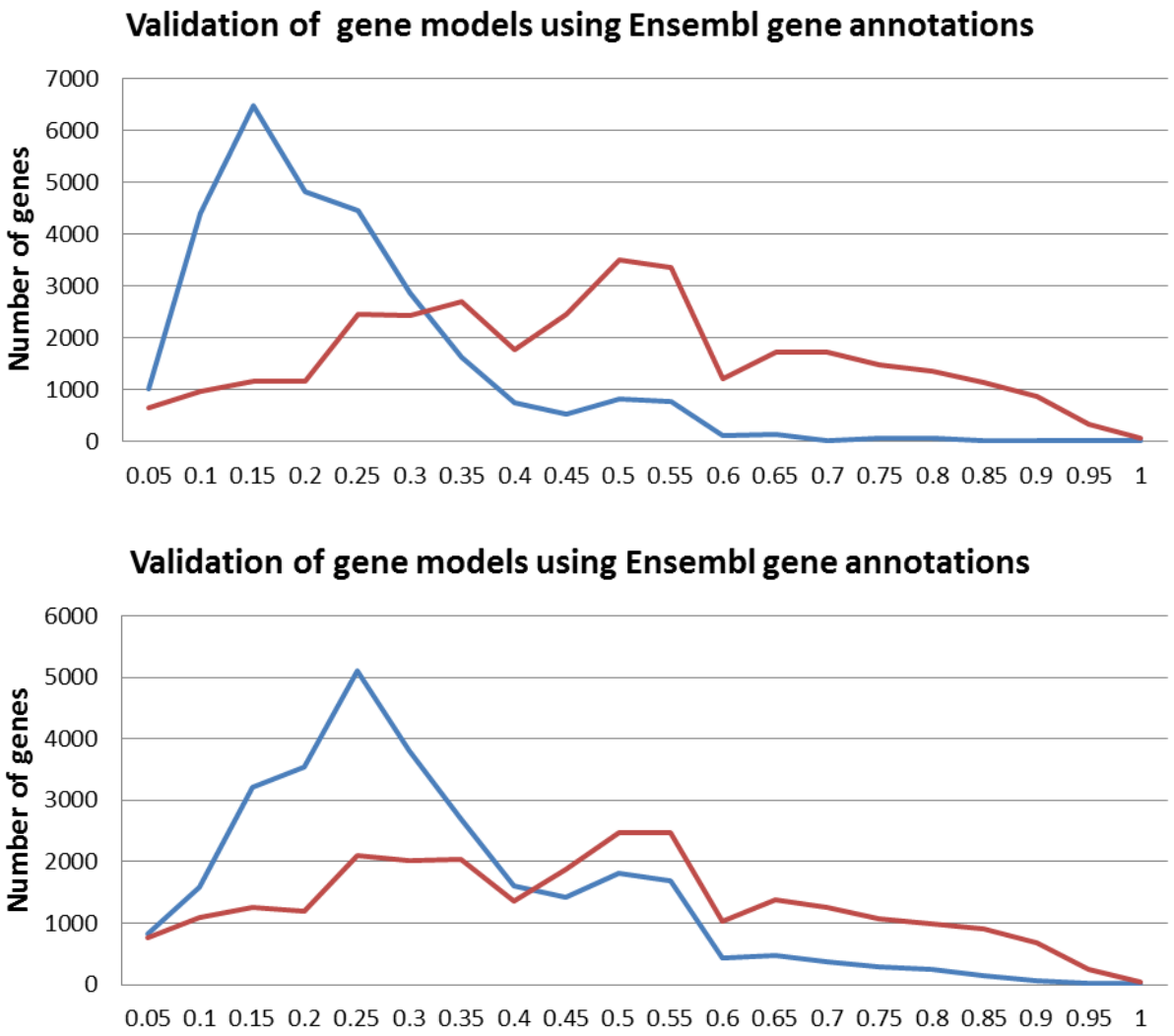


Figure 5-18. Comparisons of de novo isoform reconstructions using PASTA and Cufflinks. Blue line contains Cufflinks predictions and red line contains PASTA predictions. Comparisons are made against mouse UCSC gene annotations. These two experiments come from RNA-Seq datasets of 49 and 19 million paired end reads of 40 base pairs respectively from Mbnl3 isoform knockout experiment. The upper panel shows results from experiment of 49 million reads and lower panel shows experiment from 19 million reads. Each figure shows the number of predicted genes that are having the percentage similarity indicated by x axis compared with gene annotations

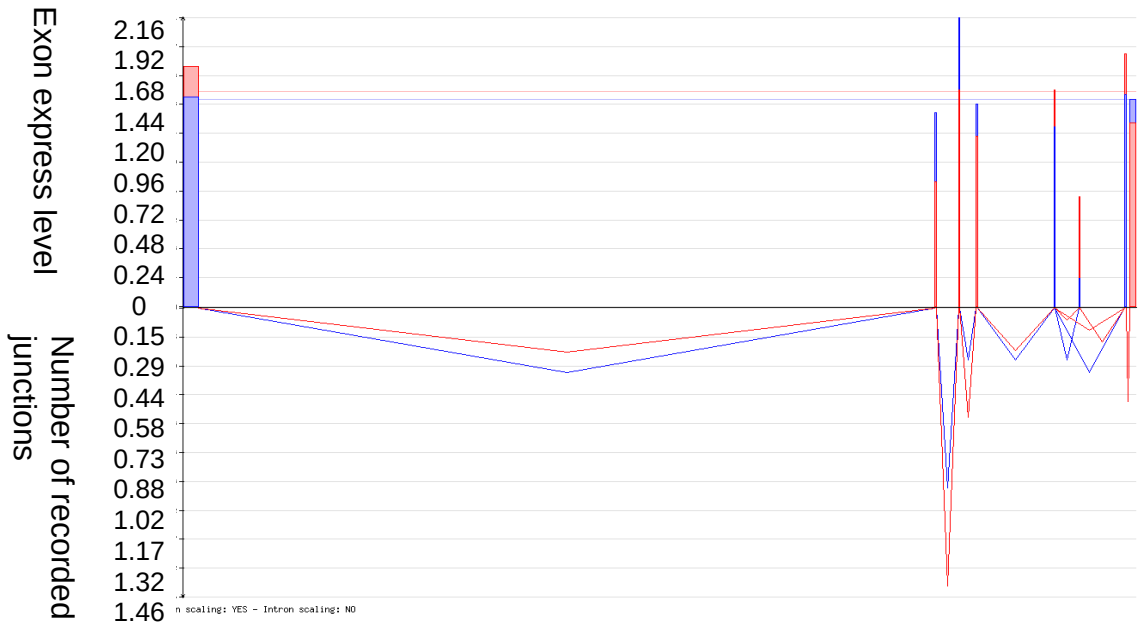


Figure 5-19. Graphical representation of gene expression using mouse gene Cnnm2 from RNA-Seq experiment (Run2). Exon and junction size are not scaled in the left panel. Introns are scaled by 20 times the logarithm of the original size. In the right panel, all intron sizes are scaled to be the equal size. Each exon or intron is displayed with two different colors with red presenting expression value from experimental group (mutant groups) and blue color from control.

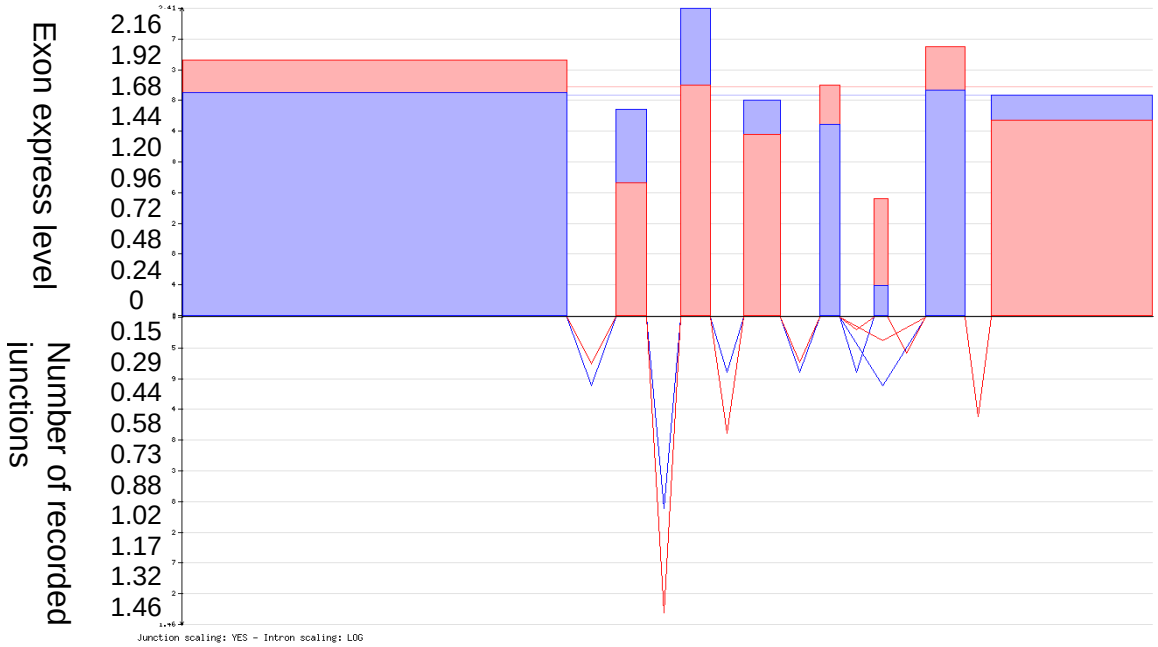


Figure 5-20. Graphical representation of gene expression using mouse gene Cnnm2 from RNA-Seq experiment (Run2) where intron size is logarithm scaled. Exon and junction size are not scaled in the left panel. Introns are scaled by 20 times the logarithm of the original size. In the right panel, all intron sizes are scaled to be the equal size. Each exon or intron is displayed with two different colors with red presenting expression value from experimental group (mutant groups) and blue color from control.

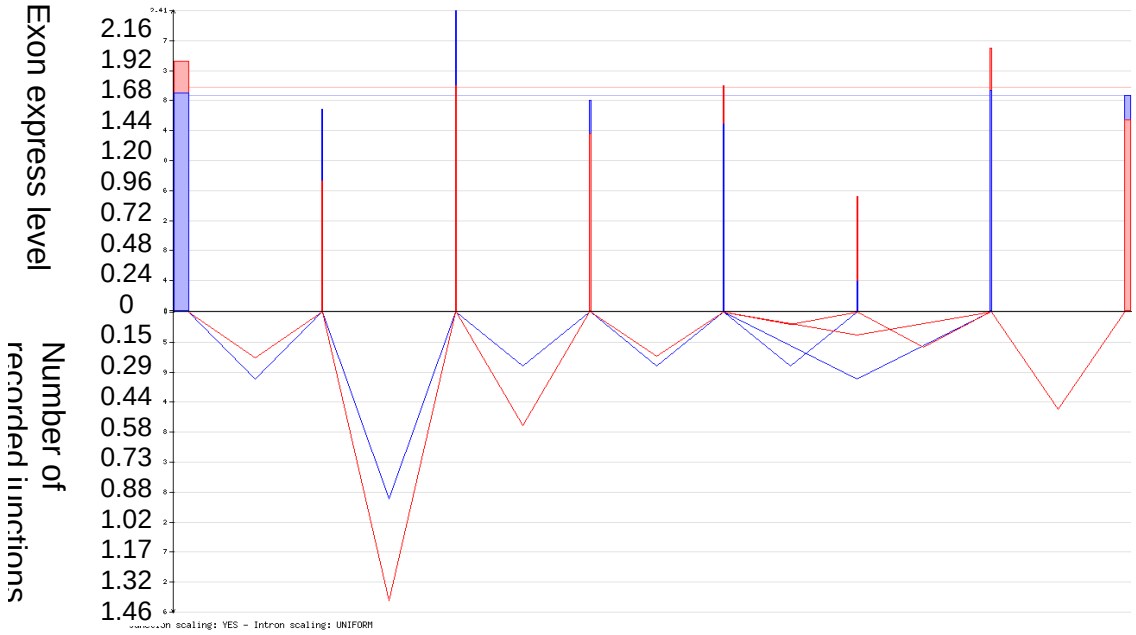


Figure 5-21. Graphical representation of gene expression using mouse gene Cnnm2 from RNA-Seq experiment (Run2) where intron is evenly spaced. Exon and junction size are not scaled in the left panel. Introns are scaled by 20 times the logarithm of the original size. In the right panel, all intron sizes are scaled to be the equal size. Each exon or intron is displayed with two different colors with red presenting expression value from experimental group (mutant groups) and blue color from control.

		Controls			Mutant				Ctrl Avg	Mut Avg
Exons	0	3.227	8.499	9.937	9.084	10.203	4.703	4.645	7.221	7.159
	1	14.203	13.972	16.738	19.025	14.551	18.278	17.692	14.971	17.386
	2	15.298	16.114	23.052	22.194	23.122	21.985	23.234	18.155	22.634
	3	15.643	15.768	13.342	21.226	17.641	13.008	15.400	14.917	16.819
	4	16.762	9.451	12.110	10.704	11.901	7.376	8.694	12.774	9.669
	5	24.094	13.737	18.656	21.129	19.824	21.233	22.508	18.829	21.173
	6	11.351	9.612	8.333	0	0	0	0	9.765	0
	7	9.108	9.233	7.843	11.305	8.262	8.812	14.101	8.728	10.620
Junctions	0_1	8.695	8.045	5.882	11.320	6.097	9.055	7.421	7.541	8.473
	1_2	14.046	4.597	9.803	18.867	6.707	12.204	8.203	9.482	11.495
	2_3	13.377	12.643	15.686	11.320	13.414	14.566	13.281	13.902	13.145
	3_4	5.685	9.195	8.823	9.433	10.365	2.755	2.343	7.901	6.224
	4_5	12.374	6.896	9.803	12.264	17.073	11.417	14.843	9.691	13.899
	5_6	11.371	3.448	6.862	0	0	0	0	7.227	0
	5_7	0.668	0	0	5.660	3.658	5.118	8.984	0.222	5.855
	6_7	0.668	0	1.960	0	0	0	0	0.876	0

Figure 5-22. Representation of Mbnl3 average exon and junction expression levels from RNA-Seq datasets of Run 2 using PASTA.

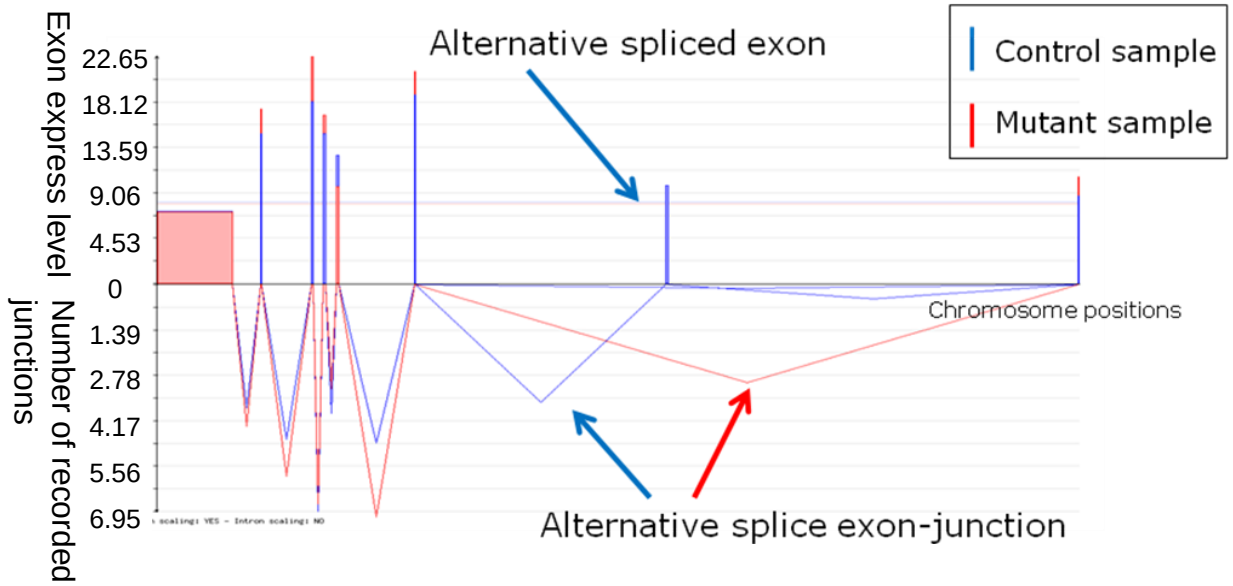


Figure 5-23. The confirmation of alternative splicing in Mbnl3 gene using Mbnl3 knockout RNA-Seq datasets. Each exon or intron is displayed with two different colors with red presenting expression value from experimental group (mutant groups) and blue color from control.

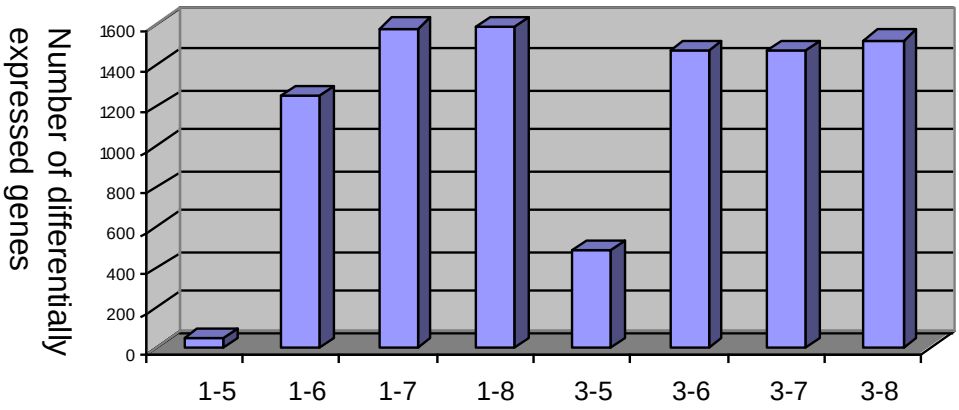


Figure 5-24. Number of differentially expressed genes by MISO using Mbnl3 RNA-Seq datasets Run2.The y-axis presents the number of alternative expressed isoforms detected by MISO.

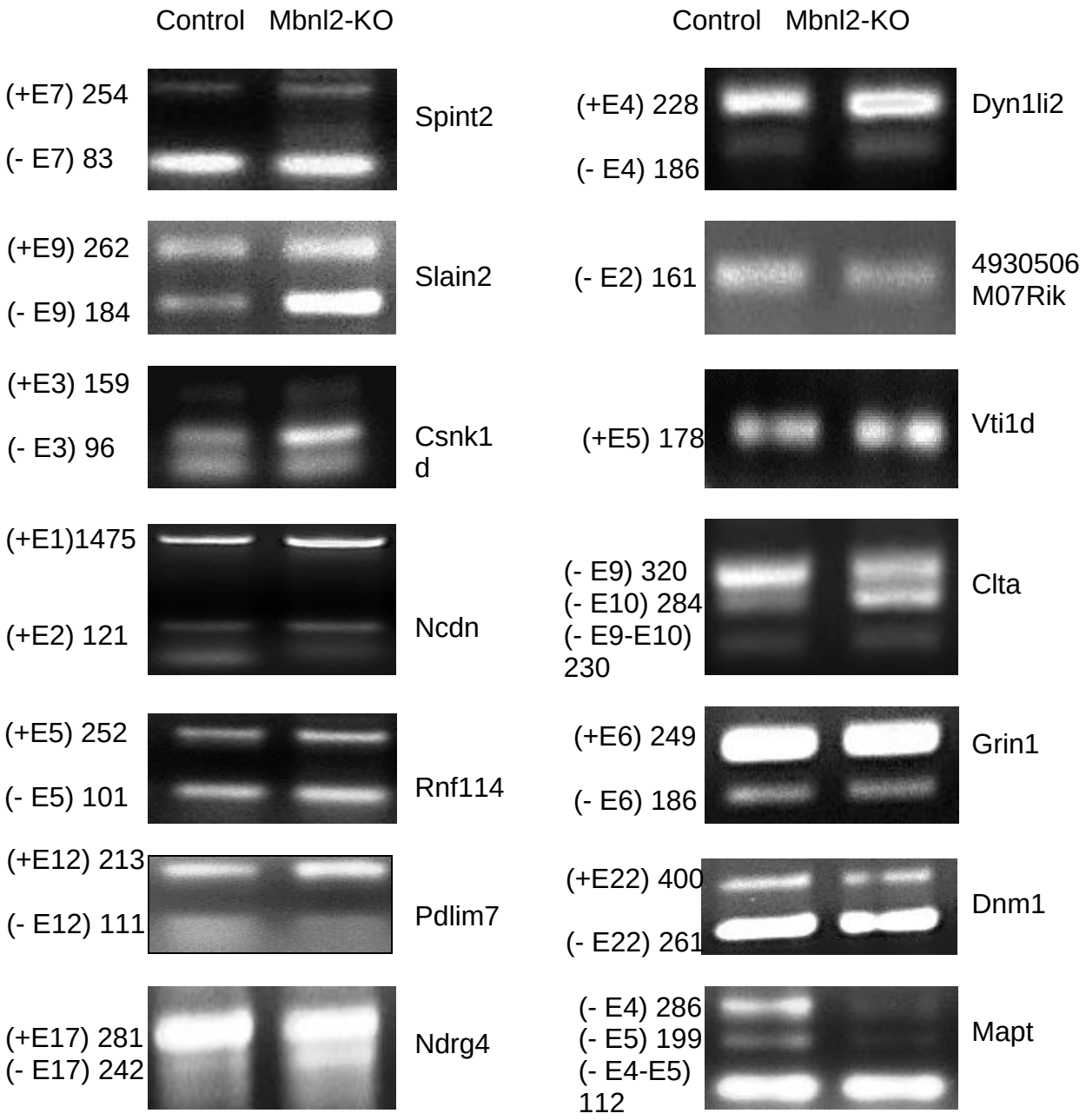


Figure 5-25. PCR results from 14 splice junction target candidates with exon skips or mutually exclusive exons. For each gene, the exon skipping event and PCR product size are labeled to the left of the PCR result, and gene name is labeled to the right of PCR result. Among these genes, Slain2, Ndrg4, Clta and Mapt are candidate genes with differentially expressed isoforms. For Ncdn, exon1 and exon2 contain alternative 5' sites. Exon2 inclusion is not detected at 4930506M07Rik. Exon5 skipping is not detected at Vti1d.

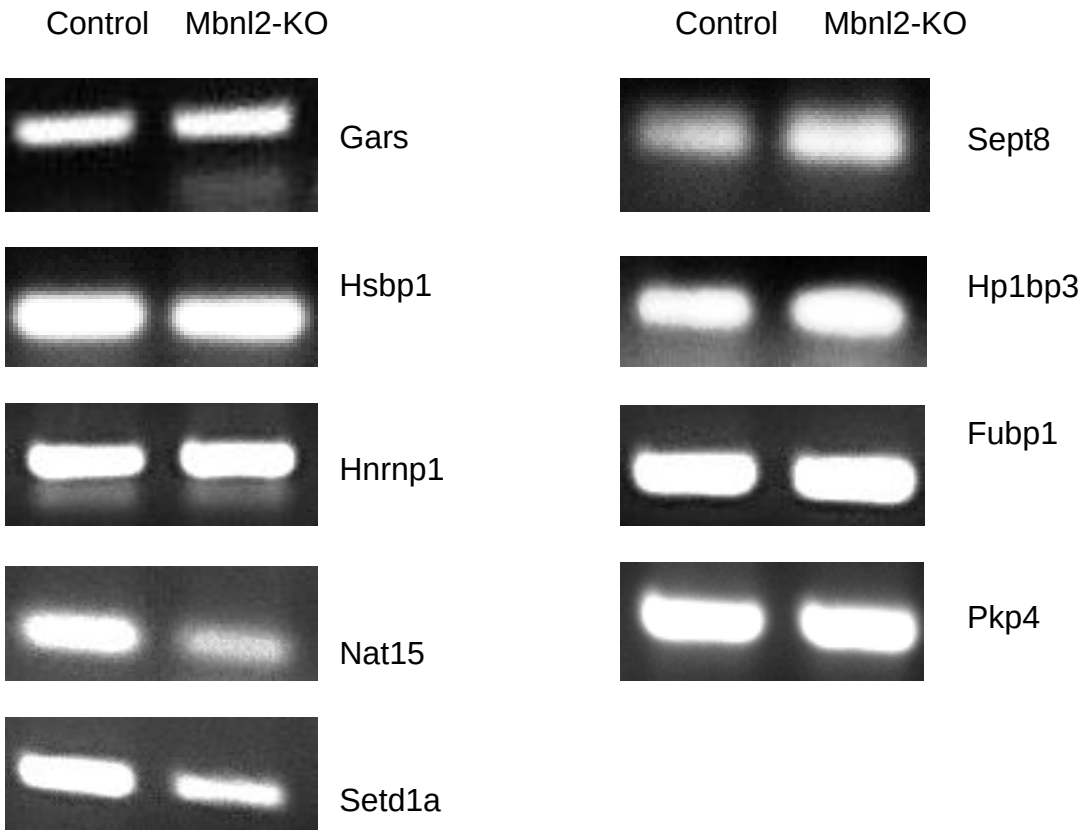


Figure 5-26. PCR results from 9 splice junction target candidates with AT-AC or GC-AG minor splice sites. Gene name is labeled to the right of each PCR result.

CHAPTER 6
CONCLUSIONS

Alternative splicing is used by a large fraction of metazoan genes to produce multiple isoforms from a single gene, thus creating transcriptome complexity and protein diversity [19]. DNA microarray technology, that has been used to profile the transcriptome, characterize genetic variation and study protein-DNA interactions for over a decade, is now being gradually replaced by the wide spread application of RNA-Seq which provides an efficient way for complete and unbiased quantification of gene expression and transcriptome composition [46, 81].

PASTA is a Large Scale Pipeline for Splicing Analysis

In this work, I have performed a comprehensive analysis of alternative splicing using PASTA on RNA-Seq datasets from mouse and maize samples. Comparisons with other popular software for RNA-Seq analysis show that PASTA can effectively identify novel splice junctions, reconstruct isoform models and estimate gene expression levels in a wide range of applications.

Splice Junction Aligner to Identify Alternative Spliced Junctions

Studies using high-throughput sequencing reveal more than 95% of human pre-mRNAs containing multiple exons generate multiple exons. Differential regulation of alternative splicing is frequently observed in different developmental stages and cellular environments [19]. Disruptions of splicing regulations have been connected with many diseases [28]. Splice junctions that span exon-exon boundaries are detected in RNA-Seq through short reads that cannot be mapped uniquely to the genome with full length matches, thus allowing for the unbiased analysis of alternative splicing and identification of novel splice junctions. A number of pipelines are proposed that identify the splice

junction from RNA-Seq datasets. PASTA is an innovative algorithm that incorporates biological context information in predicting splice junctions.

The PASTA splice junction detection module enables *de novo* splice junction identification, and is intended as the first module in a complete computational pipeline for AS analysis. Compared to similar tools, PASTA offers an increased capability to detect real splice junctions especially at low coverage levels and short sequence read size, due to several heuristic strategies it employs. It does not rely on the presence of canonical splice signals, and it uses an organism-specific statistical model to evaluate predicted intron-exon junctions. Junction positions are determined through an innovative procedure based on patterned alignments. Finally, PASTA does not filter predicted junctions on the basis of their coverage, but retains high-scoring junctions even when they are supported by a low number of reads. The reason is that the final result I am interested in is not just the presence or absence of an individual junction, but which isoform structures can be inferred from a set of junctions in the same locus. It is therefore a better strategy to retain low-coverage junctions (provided they have a high score) and evaluate the isoform (isoforms) they participate in when information about all the other junctions in them is known.

Isoform Reconstruction from RNA-Seq Datasets

Splice junction predictions enables us to reconstruct transcripts from RNA-Seq data, leading to unbiased transcriptome analysis. Identifying the whole RNA content of the transcriptome is crucial for studies on the RNA population, including alternative spliced isoforms and differentially expressed isoforms. In particular, characterization of the RNA population from RNA-Seq datasets with low throughput size and/or small read

size requires highly sensitive algorithms because of the reduced numbers of reads carrying important information about splice junctions.

Although gene annotations are available for many widely studied organisms, recent studies from high-throughput sequencing show the existence of alternative isoforms or novel transcripts that do not appear in annotations [19, 81]. Two widely used approaches for *de novo* isoform reconstruction are genome-guided and genome-independent algorithms. For species without complete reference genomes, genome-independent assembly is the only tools to identify novel transcripts [19, 81]. However, recent re-sequencing effects using high-throughput approaches have produced very high quality reference genomes for many species, thus making genome-guided transcriptome construction a better choice. RNA-Seq has been used widely to identify new transcriptional events and refine existing gene models [54, 59].

PASTA is able to reconstruct transcript from splice junction predictions and read mapping information by aligning RNA-Seq datasets to the reference genome. In comparison to Cufflinks, PASTA is able to more effectively reconstruct transcripts appearing in well-curated annotations such as ENSEMBL with closer structural similarity, even when the provided samples have a small number of reads and short read length. In addition, PASTA uses information on codon utilization preferences to increase its transcript reconstruction accuracy when prior knowledge of codon utilization is available for the organism under study.

Expression Analysis to Identify Genes with Differentially Expressed Isoforms

Many RNA-Seq platforms enable sequencing of multiple lanes in a single run. Therefore, estimating differential expression patterns from multiples lanes of RNA-Seq datasets requires adequate normalization of mapping reads and interpretation of splice

junctions, especially alternative splice junctions, and appropriate algorithms to estimate transcriptome abundances. PASTA uses expression fold changes in exon and junctions to classify genes with differentially expressed isoforms, with higher sensitivity compared with MISO and Tophat. Detailed studies on these pipelines have revealed the following distinctions. Pipelines like MISO, Cufflinks and IsoEM utilize short reads mapping to the reference genome by Tophat or Bowtie as their expression data. However, this mapping information could be biased depending on the conditions configured in the alignment, and represents short reads that are aligned to reference genome with full matches. In contrast, PASTA uses mapping information, including the same alignment information described above, and also short reads that are aligned to reference genome from either of the paired reads, and in addition, it also includes short reads that uniquely align to ends of exons with 80% -90% identity, thus reducing the information lost to the minimum extent.

PASTA also enables users to look for genes with differentially expressed isoforms using a variety of standard measures, including exon and/or junction fold changes, student t-test, K-S test and relative isoform expression levels computed by a Genetic Algorithm. In addition, it can analyze multiple control and experimental samples in the context of two sample populations in finding genes with differentially expressed isoforms. In comparison, other pipelines such as MISO rely on pairwise control and experimental comparison, introducing additional complexity on subsequent statistical analysis. PASTA can reduce the post-processing work needed for alternative splicing analysis which will facilitate analysis for the biological community. Finally, PASTA

effectively uses read mapping information and splice junction predictions, providing a more accurate evaluation of exon and/or junction expression level estimates.

Application of PASTA in Studying Splicing Regulation

The robust splicing analysis features provided by PASTA enables its application in studies on transcriptome characterization and alternative splicing regulations. To start, we can identify transcript variants, and especially alternatively spliced transcripts, using PASTA to investigate their splicing regulation in pathological conditions. Second, quantitative analysis of these transcript variants in case-control scenarios can reveal important genes and pathways that are differentially regulated. Third, PASTA enables us to find minor splicing signals, which may lead to a more in-depth understanding of minor splicing regulatory machinery.

Identification of Transcripts with Minor Splice Signals

Analysis of mouse RNA-Seq datasets using PASTA reveal the presence of minor splice signals. PCR assays confirmed the presence of all 5 GC-AG signals and 4 AT-AC signals I tested, which in turn proved PASTA's potential to accurately discover minor splice signals. However, ENSEMBL annotations only contain a total of 473 AT-AC signals, which may be an underestimation due to bias towards canonical junctions in gene-finding algorithms. PASTA identified on average 500 AT-AC signals per sequence run in mouse RNA-Seq datasets, and approximately 30-50% of these signals are present in ENSEMBL annotated junctions. The remaining 50% or more AT-AC signals may come from novel junctions. Investigation of minor splice signals can reveal important splicing regulation information since pathological conditions lead to mutations that can disrupt constitutive splicing sites, which can in turn result in the uses of minor splice sites.

Predict Splice Junctions with Small RNA-Seq Datasets and/or Small DNA Read Size

A large portion of metazoan genes are transcribed at low levels in many organisms. In addition, RNA-Seq experiments may produce a small number of short reads such as 20 million when the amount of DNA is small and a small read size of around 50bp to reduce the sequencing cost. Identifying splice junctions from these experiments require very sensitive algorithms. Results from comparison between PASTA and several other splice junction prediction tools show that PASTA is very sensitive in discovering splice junctions. Simulation studies reveal that PASTA reduced the false negative rate by more than 30% for genes transcribed at relatively low expression levels of 1 to 2 RPKM. Analysis of RNA-Seq data from *Mbnl3* and *Mbnl2* isoform knockout experiment also shows us that PASTA is able to discover more splice junctions than appear in existing gene annotations such as ENSEMBL, using a small RNA-Seq dataset of 40bp. Therefore, PASTA is especially suitable for identifying splice junctions in multiplexing experiments.

Characterization of Trans-Splicing

Another advantage of PASTA over other pipelines is its ability to characterize trans-splicing. Sequence alignment matches from distinct chromosomal locations can be effectively identified in the patterned alignment step. This enables us to study trans-splicing occurring in pre-mRNAs from different chromosomes. PASTA generates a file for trans-splicing analysis by collecting pairs of sequence fragments that originate from different chromosomes, using a model-free algorithm because of the absence of intron size information in this case.

Isoform Reconstruction for Less Studied Species

The wide application of deep sequencing technologies enables reference genomes for many less studied non-model organisms to be generated. However, existing gene annotation databases cannot provide information on complete transcriptome composition for these species. PASTA enables a reconstruction of isoform models using a small RNA-Seq dataset. Comparison with Cufflinks shows PASTA's sensitivity in reconstructing isoform models with a significantly higher similarity. In addition, PASTA can be used to evaluate the reconstructed isoform models for less studied non-model species that have a closely related and well-annotated species, through standard statistical modeling procedure using codon utilization preference. In this work, we generated isoform models *de novo* for maize and mouse and achieved a high similarity in comparison with existing annotations. The application of PASTA's isoform reconstruction module to other species will enable us to identify the transcriptome composition, thus providing invaluable information for improving gene annotation database.

PASTA is a User Friendly Pipeline

In order to reduce the computational burden related to the data analysis step, PASTA generates readable output data files that are easy to manipulate, in contrast to complex output files containing original Bayes factor in MISO. In addition, the graphical display of gene expression level from case-control studies with biological replicates provides user with direct information to study differential regulations.

Experimental Validation Confirmed PASTA's Power for Alternative Splicing Analysis

In this work, I performed a genome-wide analysis of alternative splicing using RNA-Seq datasets from *Mbnl3* and *Mbnl2* knockout datasets. The analysis reveals many new alternative splice junctions and putative new junctions, supporting a greater understanding of the transcriptome. In addition, the quantification of gene expression levels reveals genes with differentially expressed isoforms. However, the estimates vary substantially among different software, thus making it necessary to perform analysis using different platforms and selecting the targets with careful consideration on the software's underlying assumptions and algorithms.

Outlook

I am planning to extend the software package developed in this work to detect nontraditional splicing patterns such as circular RNAs [82]. In addition, I will improve the algorithm to work on longer DNA short reads from RNA-Seq to accommodate for detection of two or more splice junctions out of a single DNA short read.

APPENDIX
SEQUENCING PREPARATION AND SOFTWARE DEVELOPMENT PLATFORMS

Reference Genomes Used for Model Organisms under Study

Mouse reference genome NCBI37, human reference genome build 18 and maize reference genome build z37 were used as standard reference genome. These reference genomes are downloaded from the UCSC reference genome database.

Online RNA-Seq Datasets from NCBI Short Read Archie (SRA)

Three RNA-Seq datasets from Illumina platform were downloaded from the NCBI SRA database.

- Illumina sequencing of mouse brain transcript fragment library (accession number “SRX000350”).
- Illumina sequencing of mouse liver transcript fragment library (accession number “SRX000351”).
- Illumina sequencing of mouse muscle transcript fragment library (accession number “SRX000352”).

RNA-Sequencing Platform

Samples were sequenced via Illumina Genome Analyzer IIx kindly provided by Department of Molecular Genetics and Microbiology at the University of Florida. The RNA-Seq cDNA library preparation is performed using the library preparation key described in Table 6-1.

RNA-Seq Experiments using Mouse Samples

The *Mbnl3* isoform knockout mice were kindly provided by Dr Swanson’s lab. The 2nd exon from *Mbnl3* is knockout to produce an alternative isoform resulting in a truncated protein of 27 kd. The final polyA mRNA from muscle of control and isoform knockout mice are used for cDNA library construction and RNA sequencing.

Mbnl3 Isoform Knockout Design in Mouse

Mbnl3 is located in chromosome 14. It has 8 exons, and it is transcribed from the minus strand. *Mbnl3* isoform knockout is generated by knocking down the second exon to generate an alternative *Mbnl3* isoform with absence of exon 2. The forelimb tissues are collected from mouse of embryo stage 14.5E.

Mbnl3 isoform knockout mice RNA-Seq run1

The experiment was conducted on February, 2011 with paired-end RNA-Seq datasets of read size of 40 nucleotides. RNA-Seq reads are generated from 2 control lanes and 4 mutant lanes.

Mbnl3 isoform knockout mice RNA-Seq run2

The experiment was conducted on December, 2010 with paired-end RNA-Seq datasets of read size of 51 nucleotides. RNA-Seq reads are generated from 3 control lanes and 4 mutant lanes.

Mbnl2 Isoform Knockout Design in Mouse

Mbnl2 mouse gene is located in chromosome X. This experiment is conducted on March, 2011 and produced paired-end RNA-Seq reads of 40bp.

RNA-Seq Experiments using Maize Samples

Three RNA-Seq datasets are kindly provided by Brad W Barbazuik.

- Control: “FSU" obtained from Illumina RAN-Seq platform with single end sequencing of read size of 76 nucleotides.
- Mutant: "Sample3” and “Sample4” respectively from SOLiD sequencing platforms with single end sequencing of read size of 60 nucleotides.

Software Implementation Environment

System Requirements

I used the Linux operating system as our major development environment.

Data Processing System

The system I am using for most of our testing and data processing work is in a standard Linux system with 24 cores with 64 GB of RAM.

The Choice of a Programming Language

The choice of the right programming language not only should allow accurate implementation, but it should be ready to be extended for years. LISP is a programming language that is used widely in artificial intelligence, and it is equally suitable for this research where I am extracting gene features from many short read datasets. In this work, LISP is used for the major language for implementing PASTA. In addition, part of the implementation is also including bash shell scripts.

Software Implementation Tool

I used the GNU Emacs as the development environment for software implementation and debugging.

Relational Database

In this work, I used SQL which is the world's most used relational database management system as the major platform for important storages of gene annotations. SQL offers a fast, reliable way to manipulate data such as insert, delete and update of data records. SQL has many important data manipulation properties such as efficient and concurrent data access, easy-to-use query language that facilitate efficient data retrieval. SQL database are construct in the form of tables, where user could define a list of attributes such as gene identifier, transcription start site and other information.

Tables from gene annotations of several sources, including RefSeq and ENSEMBL, are used to build the relation database for the study. A row of data is added to the table whenever the user insert a list of records specified by the defined attributes. SQL is very effective to define relation schemas that are used to associate the tables in a specific mapping relationship. For example, a one-to-one mapping associates the two tables so that a data row of a certain attribute is matched by another data row of the second table of the same attribute.

Table A-1. Flowchart of cDNA library preparation protocol

Step	Preparation steps	Stage
1	Extract total RNA using trizol protocol	mRNA preparation
2	Anneal adaptors together	
3	mRNA purification from 75ug total RNA using oligo dT Dynabeads (Invitogen)	
4	Fragmentation of mRNA	
5	First strand cDNA synthesis	Synthesis
6	Second Strand cDNA Synthesis	
7	End Repair (Epicentre End-IT Repair Kit)	Adaptation
8	Addition of a single <A> base (NEB)	
9	Adaptor ligation (Epicenter Fast-Link DNA Ligation Kit)	
10	Gel Purification of cDNA templates (Zymo Research)	Enrichment
11	PCR Enrichment of Purified cDNA Templates	

Table A-2. The coordinates of Mbnl3 gene in mouse genome build mm9.

Original exon coordinates of Mbnl3 gene			Exon coordinates of Mbnl3 isoform knockout		
Exon ID	Exon Start	Exon End	Exon ID	Exon Start	Exon End
1	48474187	48474195	1	48474187	48474195
2	48476994	48477088	2	48476994	48477088
3	48482098	48482248	3	48482098	48482248
4	48483282	48483518	4	48483282	48483518
5	48484606	48484797	5	48484606	48484797
6	48492381	48492545	6	48492381	48492434
7	48517595	48517771			

Table A-3. The translated amino acid sequences of Mbnl3 gene (full length) and Mbnl3 isoform knockout (in bold).

Protein Sequences
M T P V N V A L I R D T K W L T L E V C R E F Q R G T C S R A D A E C R F A H P P R V C H V E N G R V V A C F D S L K G R C T R E N C K Y L H P P P H L K S Q L E V N G R N N L I Q Q K T A A A M F A Q H M Q L M L Q N A Q M S S L A S F P M N P S L A A N P A M A **F N P Y M T H P G M G L V P A E L L P N G P V L I S G N P P L A L P G V P G P K P I R T D R L E V C R E F Q R G N C T R G E S E C R Y A H P T D V S M I E V T D N S V T I C M D Y I K G R C S R E K C K Y F H P P P H L Q A K L R A A H H Q M N H S A A N A M A L P H G A L Q L I P K R S A L D K A N G A T P V F N P S V F H C Q Q A L A N M Q I P Q Q A F I P T V P M M H G A T P S T V S T A T P P A S N V P Y V P T T T G N Q L K Y ***

LIST OF REFERENCES

1. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annu. Rev. Biochem.* 2003, **72**:291–336.

2. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**:561–563.

3. Roy H, Ibba M: **Molecular biology: Sticky end in protein synthesis**. *Nature* 2006, **443**:41–42.

4. Chamberlin M: **Transcription 1970: A Summary**. *Cold Spring Harb Symp Quant Biol* 1970, **35**:851–873.

5. Koonin EV: **Does the central dogma still stand?** *Biol. Direct* 2012, **7**:27.

6. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA**. *Proc. Natl. Acad. Sci. U.S.A.* 1977, **74**:3171–3175.

7. Cheng SC, Abelson J: **Spliceosome assembly in yeast.** *Genes Dev.* 1987, **1**:1014–1027.

8. Konarska MM, Grabowski PJ, Padgett RA, Sharp PA: **Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors**. *Nature* 1985, **313**:552–557.

9. Matlin AJ, Clark F, Smith CWJ: **Understanding alternative splicing: towards a cellular code**. *Nat. Rev. Mol. Cell Biol.* 2005, **6**:386–398.

10. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity**. *Nature Genetics* 2004, **36**:1255–1257.

11. Burset M, Seledtsov IA, Solovyev VV: **Analysis of canonical and non-canonical splice sites in mammalian genomes**. *Nucleic Acids Res.* 2000, **28**:4364–4375.

12. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB: **Accumulation of GC donor splice signals in mammals**. *Biol. Direct* 2008, **3**:30.

13. Farrer T, Roller AB, Kent WJ, Zahler AM: **Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing**. *Nucleic Acids Res.* 2002, **30**:3360–3367.

14. Hodge MR, Cumsky MG: **Splicing of a yeast intron containing an unusual 5' junction sequence.** *Mol. Cell. Biol.* 1989, **9**:2765–2770.

15. Cooper TA, Wan L, Dreyfuss G: **RNA and disease**. *Cell* 2009, **136**:777–793.

16. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: **Deciphering the splicing code**. *Nature* , **465**:53–59.

17. Schnable PS, Ware D, Fulton RS, et al.: **The B73 Maize Genome: Complexity, Diversity, and Dynamics**. *Science* 2009, **326**:1112–1115.

18. Barbazuk WB, Fu Y, McGinnis KM: **Genome-wide analyses of alternative splicing in plants: opportunities and challenges**. *Genome Res.* 2008, **18**:1381–1392.

19. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, **463**:457–463.

20. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**:470–476.

21. Änkö M-L, Müller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM: **The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes**. *Genome Biology* 2012, **13**:R17.

22. Kim E, Goren A, Ast G: **Alternative splicing: current perspectives**. *Bioessays* 2008, **30**:38–47.

23. Brogna S, Wen J: **Nonsense-mediated mRNA decay (NMD) mechanisms**. *Nature Structural & Molecular Biology* 2009, **16**:107–113.

24. Cole SE, LaRiviere FJ, Merrikh CN, Moore MJ: **A Convergence of rRNA and mRNA Quality Control Pathways Revealed by Mechanistic Analysis of Nonfunctional rRNA Decay**. *Molecular Cell* 2009, **34**:440–450.

25. Maquat LE: **Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics**. *Nature Reviews Molecular Cell Biology* 2004, **5**:89–99.

26. Chang Y-F, Imam JS, Wilkinson MF: **The nonsense-mediated decay RNA surveillance pathway**. *Annu. Rev. Biochem.* 2007, **76**:51–74.

27. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E, Lucier J-F, Thibault P, Rancourt C, Tremblay K, Prinos P, Chabot B, Elela SA: **Cancer-associated regulation of alternative splicing**. *Nature Structural & Molecular Biology* 2009, **16**:670–676.

28. Wang G-S, Cooper TA: **Splicing in disease: disruption of the splicing code and the decoding machinery**. *Nature Reviews Genetics* 2007, **8**:749–761.

29. Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy**. *Nature Biotechnology* 2004, **22**:535–546.

30. Krawczak M, Reiss J, Cooper DN: **The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences**. *Hum. Genet.* 1992, **90**:41–54.

31. Xing Y, Lee C: **Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes**. *Nature Reviews Genetics* 2006, **7**:499–509.

32. Steinbach P, Gläser D, Vogel W, Wolf M, Schwemmle S: **The DMPK gene of severely affected myotonic dystrophy patients is hypermethylated proximal to the largely expanded CTG repeat.** *Am J Hum Genet* 1998, **62**:278–285.

33. Kanadia RN, Urbinati CR, Crusselle VJ, Luo D, Lee Y-J, Harrison JK, Oh SP, Swanson MS: **Developmental expression of mouse muscleblind genes Mbnl1, Mbnl2 and Mbnl3**. *Gene Expr. Patterns* 2003, **3**:459–462.

34. Lee K-S, Smith K, Amieux PS, Wang EH: **MBNL3/CHCR prevents myogenic differentiation by inhibiting MyoD-dependent gene transcription**. *Differentiation* 2008, **76**:299–309.

35. Squillace RM, Chenault DM, Wang EH: **Inhibition of muscle differentiation by the novel muscleblind-related protein CHCR**. *Dev. Biol.* 2002, **250**:218–230.

36. Miller JW, Urbinati CR, Teng-umnuay P, Stenberg MG, Byrne BJ, Thornton CA, Swanson MS: **Recruitment of human muscleblind proteins to (CUG)n expansions associated with myotonic dystrophy**. *The EMBO Journal* 2000, **19**:4439–4448.

37. Licatalosi DD, Darnell RB: **RNA processing and its regulation: global insights into biological networks**. *Nature Reviews Genetics* 2010, **11**:75–87.

38. Salomonis N, Nelson B, Vranizan K, Pico AR, Hanspers K, Kuchinsky A, Ta L, Mercola M, Conklin BR: **Alternative Splicing in the Differentiation of Human Embryonic Stem Cells into Cardiac Precursors**. *PLoS Comput Biol* 2009, **5**.

39. Li Q, Lee J-A, Black DL: **Neuronal regulation of alternative pre-mRNA splicing**. *Nature Reviews Neuroscience* 2007, **8**:819–831.

40. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nat. Methods* 2010, **7**:1009–1015.

41. Twine NA, Janitz K, Wilkins MR, Janitz M: **Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease**. *PLoS ONE* 2011, **6**:e16266.

42. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10**:57–63.

43. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol.* 2009, **10**:R25.

44. Trapnell C, Pachter L, Salzberg SL: **TopHat: Discovering Splice Junctions with RNA-Seq**. *Bioinformatics* 2009, **25**:1105–1111.

45. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I: **De novo assembly and analysis of RNA-seq data**. *Nature Methods* 2010, **7**:909–912.

46. Shendure J: **The beginning of the end for microarrays?** *Nat. Methods* 2008, **5**:585–587.

47. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res.* 2008, **18**:1509–1517.

48. Harbig J, Sprinkle R, Enkemann SA: **A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array**. *Nucl. Acids Res.* 2005, **33**:e31–e31.

49. Cloonan N, Grimmond SM: **Transcriptome content and dynamics at single-nucleotide resolution**. *Genome Biology* 2008, **9**:234.

50. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M-L: **A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome**. *Science* 2008, **321**:956–960.

51. Lamm AT, Stadler MR, Zhang H, Gent JI, Fire AZ: **Multimodal RNA-seq using single-strand, double-strand, and CircLigase-based capture yields a refined and extended description of the C. elegans transcriptome**. *Genome Res.* 2011, **21**:265–275.

52. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat. Genet* 2008, **40**:1413–1415.

53. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, Haas SA, Yaspo M-L: **Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments**. *Nucl. Acids Res.* 2010, **38**:e112–e112.

54. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature Methods* 2008, **5**:621–628.

55. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ van, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq**

reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 2010, **28**:511–515.

56. Dilworth RP: **A Decomposition Theorem for Partially Ordered Sets**. *The Annals of Mathematics* 1950, **51**:161.

57. Reed R, Maniatis T: **The role of the mammalian branchpoint sequence in pre-mRNA splicing**. *Genes Dev.* 1988, **2**:1268–1276.

58. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: **Annotating non-coding regions of the genome**. *Nature Reviews Genetics* 2010, **11**:559–571.

59. Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F: **Annotating genomes with massive-scale RNA sequencing**. *Genome Biol.* 2008, **9**:R175.

60. Burge CB, Karlin S: **Finding the genes in genomic DNA**. *Current Opinion in Structural Biology* 1998, **8**:346–354.

61. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A: **Codon Catalog Usage and the Genome Hypothesis**. *Nucl. Acids Res.* 1980, **8**:197–197.

62. Ikemura T: **Codon Usage and tRNA Content in Unicellular and Multicellular Organisms.** *Mol Biol Evol* 1985, **2**:13–34.

63. Burghes AHM, Beattie CE: **Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick?** *Nature Reviews Neuroscience* 2009, **10**:597–609.

64. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou Y-C, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJM, Tai IT, Marra MA: **Alternative expression analysis by RNA sequencing**. *Nature Methods* 2010, **7**:843–847.

65. Holt I, Jacquemin V, Fardaei M, Sewry CA, Butler-Browne GS, Furling D, Brook JD, Morris GE: **Muscleblind-like proteins: similarities and differences in normal and myotonic dystrophy muscle**. *Am. J. Pathol.* 2009, **174**:216–227.

66. Teplova M, Patel DJ: **Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1**. *Nat. Struct. Mol. Biol.* 2008, **15**:1343–1351.

67. Charizanis K, Lee K-Y, Batra R, Goodwin M, Zhang C, Yuan Y, Shiue L, Cline M, Scotti MM, Xia G, Kumar A, Ashizawa T, Clark HB, Kimura T, Takahashi MP, Fujimura H, Jinnai K, Yoshikawa H, Gomes-Pereira M, Gourdon G, Sakai N, Nishino S, Foster TC, Ares Jr. M, Darnell RB, Swanson MS: **Muscleblind-like 2-Mediated Alternative**

Splicing in the Developing Brain and Dysregulation in Myotonic Dystrophy. *Neuron* 2012, **75**:437–450.

68. Lee K-S, Cao Y, Witwicka HE, Tom S, Tapscott SJ, Wang EH: **RNA-binding Protein Muscleblind-like 3 (MBNL3) Disrupts Myocyte Enhancer Factor 2 (Mef2) β-Exon Splicing**. *J. Biol. Chem.* 2010, **285**:33779–33787.

69. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.

70. Aban IB, Meerschaert MM, Panorska AK: **Parameter Estimation for the Truncated Pareto Distribution**. *Journal of the American Statistical Association* 2006, **101**:270–277.

71. Gao K, Masuda A, Matsuura T, Ohno K: **Human branch point consensus sequence is yUnAy**. *Nucleic Acids Res.* 2008, **36**:2257–2267.

72. Sickmier EA, Frato KE, Shen H, Paranawithana SR, Green MR, Kielkopf CL: **Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65**. *Mol. Cell* 2006, **23**:49–59.

73. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives**. *Nucleic Acids Res* 2009, **37**:D32–D36.

74. Smith TF, Waterman MS, Fitch WM: **Comparative biosequence metrics**. *Journal of Molecular Evolution* 1981, **18**:38–46.

75. Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW: **An assessment of gene prediction accuracy in large DNA sequences**. *Genome Res.* 2000, **10**:1631–1642.

76. Fickett JW, Tung CS: **Assessment of protein coding measures**. *Nucleic Acids Res.* 1992, **20**:6441–6450.

77. Burset M, Guigó R: **Evaluation of gene structure prediction programs**. *Genomics* 1996, **34**:353–367.

78. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)**. *Bioinformatics* 2011, **27**:2518–2528.

79. Hull J, Campino S, Rowlands K, Chan M-S, Copley RR, Taylor MS, Rockett K, Elvidge G, Keating B, Knight J, Kwiatkowski D: **Identification of common genetic variation that modulates alternative splicing**. *PLoS Genet.* 2007, **3**:e99.

80. Nicolae M, Mangul S, Măndoiu I, Zelikovsky A: **Estimation of alternative splicing isoform frequencies from RNA-Seq data**. *Algorithms for Molecular Biology* 2011, **6**:9.

81. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nat. Methods* 2011, **8**:469–477.

82. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO: **Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types**. *PLoS ONE* 2012, **7**:e30733.

BIOGRAPHICAL SKETCH

Shaojun Tang earned his Bachelor of Science degree in biological science from Sun Yat-Sen University in 2006. He has been the recipient of many awarded during his study in Sun Yat-sen University, and he is one of the only three recipients to do a research internship in National University of Singapore at 2006. While he is working in College of Medicine at University of Florida since 2007, he is also pursuing a Master of Science degree in computer engineering.