Stony Brook University

Novel Computational Methodology for Detecting and Quantifying Alternative Splicing from RNA-Seq data

A Dissertation Presented

by

Jie Wu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

August 2013

Stony Brook University

The Graduate School

Jie Wu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Michael Q. Zhang – Dissertation Advisor
Adjunct Professor, Applied Mathematics and Statistics, Stony Brook University and Cold Spring Harbor Laboratory
Cecil H. and Ida Green Distinguished Chair Professor
Department of Molecular and Cell Biology, the University of Texas at Dallas

Haipeng Xing - Chairperson of Defense
Associate Professor, Applied Mathematics and Statistics

Wei Zhu
Professor, Applied Mathematics and Statistics

Adrian R. Krainer
Professor, Cold Spring Harbor Laboratory

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Novel Computational Methodology for Detecting and Quantifying Alternative Splicing from RNA-Seq data

by

Jie Wu

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

2013

Recent development of ultra-high-throughput sequencing of the transcriptome (mRNA-Seq) provides a means of profiling RNA splicing events at unprecedented depth. On the other hand, the ultra-high coverage and the complexity brought by mRNA-Seq data also create big challenges for computational analysis. My Ph.D. work focuses on developing algorithms to detect, quantify and characterize alternative splicing (AS) from mRNA-Seq data. These algorithms include:

(1) OLego, a fast and sensitive splice mapping program for mRNA-Seq data. The most important features of OLego include strategic and efficient searches with very small seeds (12~14 nt), and a built-in regression model to score exon junctions. In addition, OLego does not require any external mapper, and is implemented in C++ with full support of

multithreading. As a consequence, OLego has improved sensitivity on junction and exon discovery while keeping high accuracy and speed.

(2) In-house scripts to identify AS events from alignment results of mRNA-Seq data. Instead of constructing full structures of the transcripts, this approach identifies exons and AS events from the junction reads directly to achieve lower complexity and higher sensitivity of splicing events.

(3) SpliceTrap, a method to quantify exon inclusion ratios from paired end mRNA-Seq data using a Bayesian model. The algorithm solves the splicing problem by looking at local splicing events instead of the whole transcripts, which enables quantification of exon inclusion ratios without knowing the complete transcript structure. It also utilizes prior information including fragment size distribution and inclusion ratio models from highly covered AS events.

All of the programs above are splicing-centric tools and can be used to study AS events with high resolution and sensitivity. We have applied this pipeline on many real dataset including the BodyMap 2.0 data, in which we identified 120,110 cassette exons in human genome, including 82,528 novel cassette exon events. Strikingly, we identified over 2,000 cassette micro-exons smaller than 27 nt, 105 of them have a length of 6 nt. Because of the minimal information that can be possibly encoded in this set of exons, they serve as an excellent model to study their functional significance and mechanism of AS regulation.

Dedicated to my family

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA: Alternative Accepter

AD: Alternative Donor

AS: Alternative Splicing

BWT: Burrow-Wheeler Transform

bp: base pair

CA: CAssette exons

CAD: Cassette Exon Discovery rate

CAGE: Cap Analysis of Gene Expression

CLIP-Seq: Cross-Linking ImmunoPrecipitation-high-throughput Sequencing

CS: ConStitutive exons

CSD: ConStitutive exon Discovery rate

dc: dynamic cutoff

EJC: Exon junction complex

EM: Expectation-Maximization

ESE: Exonic Splicing Enhancer

ESS: Exonic Splicing Silencers

EST: Expressed Sequence Tag

FDR: False Discovery Rate

FNR: False Negative Rate

FSM: Fragment Size Model

FPKM: Fragments Per Kilobase of transcript per Million mapped reads

GMM: Gaussian Mixture Model

IR: Intron Retention

ir: inclusion ratio

IRM: Inclusion Ratio Model

ISE: Intronic Splicing Enhancer

ISS: Intronic Splicing Silencers

LASSO: Least Absolute Shrinkage and Selection Operator

MLE: Maximum Likelihood Estimation

MPSS: Massively Parallel Signature Sequencing

nt: nucleotide

NGS: Next Generation Sequencing

NMD: Nonsense-Mediated Decay

ORF: Open Reading Frame

PCC: Pearson Correlation Coefficient

PPV: Positive Predictive Value

PTC: Pre-mature stop (Termination) Codon

RBP: RNA Binding Protein

RPKM: Reads Per Kilobase of transcript per Million mapped reads

RT-PCR: Real-Time Polymerase Chain Reaction

RUST: Regulated Unproductive Splicing and Translation

SAGE: Serial Analysis of Gene Expression

SN: Sensitivity

SP: Specificity

Acknowledgments

First, I would gratefully thank my advisor Dr. Michael Q. Zhang for his support and advice during my Ph.D. study. He led me into the field of genomic bioinformatics and always has constructive comments on my projects. I would like to deeply thank Dr. Adrian R. Krainer. I learned a lot about biology in his lab and I had many useful discussions in his lab meetings. He also provided office space for closer interaction with biologists in his lab. His lab also provides the experimental support for my projects. I would like to thank Dr. Haipeng Xing, who provides many useful discussions on statistical models in my work. I acknowledge Dr. Wei Zhu for her support as my thesis committee member. I would like to deeply thank Dr. Chaolin Zhang. He is a wonderful computational biologist and I had many useful discussions with him, and he advised me on the OLego project. I especially thank Martin Akerman, who is a supportive friend and he helps me on all of my projects throughout my Ph.D. work. I learned a lot from him. I gratefully thank my labmates in both Zhang Lab and Krainer Lab, Yifan Mo helped me a lot on statistics, Will Liao had many useful discussions with me in the lab. Olga Anczuków helped to carry out PCR validations. I also acknowledge other former and current members and visiting scientists of Zhang Lab (Xiaotu Ma, Zhenyu Xuan, Chenghai Xue, Xiaowo Wang, Yunfei Wang, Pradipta Ray, Wen-Yu Chung, Zhihua Zhang, Justin Kinney, Yongjin Li, Lirong Zhang, Weijun Luo, Man-Huang Eric Tang, Kazimierz Wrzeszczynski, Darrin Lewis, Mamoru Kato, Martin Paradesi, Nevenka Dimitrova, Vinay Varadan, Xiaoping Zhao, Ashwinikumar Kulkarni, Weilong Guo, Yu Liu, Zefeng Zhang and Hao Lin) and Krainer Lab (Shuying Sun, Xavier Roca, Lisandra Zepeda, Yimin Hua, Kuan-Ting Lin, Yilei Liu, Rahul Sinha, Kentaro Sahashi, Isabel Aznarez, Shipra Das, Ying Hsiu Liu, Jaclyn Novatt, Fatma Bezirci, Hyun-Yong Jeon, Tomoki Nomakuchi, Deblina Chatterjee, Ruei-Ying Tzeng, Mads Jensen, Oliver Fregoso and Zhenxun Wang) for their comments and discussions on my projects. I want to thank my classmates, colleagues and friends in Stony Brook and Cold Spring Harbor.

Finally, I deeply thank my wife Qinghong Yan, who always supports me and stays with me all the time. I am the luckiest person with her accompany. I also gratefully thank my parents Qiwen Wu and Lin Li, they always understood and supported whatever decisions I made and encouraged me to pursue my career.

Chapter 1 :

Introduction

In the year of 1953, James Watson and Francis Crick proposed the double helix structure of deoxyribonucleic acids (DNA)(1). And four years later, Francis Crick stated the theory of the central dogma of molecular biology, in which the sequential information flows from DNA to RNA and then protein (2). For decades since then, researchers have been working hard to understand the mechanism behind the central dogma. Most recently, along with the completion of human genome project (3), people started to focus more on the -omics study, using rapidly evolving sequencing techniques. In this Chapter, I will briefly describe the biological background and the techniques related to my Ph.D. study, including brief introduction of RNA splicing, next generation sequencing (NGS) technique and associated bioinformatics methods.

The central dogma of molecular biology and RNA splicing

The sequence information encoded in the genome, or DNA, is first transferred to RNA (transcription), which is later translated to protein (translation). It is also known that the information can be copied from DNA to DNA, or RNA to RNA, in the process of DNA or RNA replication, and flow backwards from RNA to DNA in reverse transcription. This is so called “the central dogma of molecular biology”. In this dissertation, I only focus on part of the process of transcription.

In the transcription of DNA, segments in DNA are read by RNA polymerases and complementary RNA strands are synthesized. In eukaryotic cells, the RNA must be processed with further steps, including 5' capping, polyadenylation and splicing. In the RNA splicing step, the introns are spliced out and the exons are joined to form the final mature messenger RNAs (mRNA), which contain coding information for protein synthesis.

Mechanism of RNA splicing

The RNA splicing was first discovered in the 1970's (4). Through these decades, the mechanism of splicing has been studied in many organisms and is now relatively well characterized. The introns are removed by cleavage at conserved sequences at both ends of the intron, which are called 5' and 3' splice sites. It is well known that these sites consist of conserved di-nucleotide sequences (GU at the 5' splice site and AG at the 3' splice site). And other conserved sequence found in the intron is the "branch point" located 18-40 nucleotide (nt) upstream of the 3' splice site, which always has an adenine (A) in the middle (5).

RNA splicing is coordinated by a complex consist of RNAs and proteins (spliceosome)(6), In the beginning of the splicing process, a small nuclear ribonucleoproteins (snRNPs) called U1 attaches to the 5' splice site, followed by the cleavage of the pre-mRNA at this position. This cut end then binds to the branch point in the intron and forms a loop called lariat, Afterwards, in the coordination with other snRNPs (U2, U4/U6 and U5), the two flanking exons are joined with covalent bond and the lariat is released (5).

Alternative splicing and different splice patterns

The splice pattern can be altered in different ways (5,7). In most cases, the pattern determines whether a portion of sequence region included or excluded from the final transcript, thus affecting the amino acid sequence of the translated proteins. This phenomenon, so called alternative splicing (AS), allows the genome to encode more proteins with limited number of protein-coding genes. For example, it has been reported

that more than 90% of human genes are undergoing AS, which largely increase the protein diversity in human (8,9).

The most common AS in eukaryotic is cassette exon (7). While some of the exons are constitutively spliced in the RNA splicing process and included in the final mRNA (constitutive exons), there are some exons which are regulated and can be either included or excluded in the final transcript. More than 40% of the AS events can be categorized into this group (7,10). There are also other types of AS events. Sometimes, the exons have variant lengths by using different 5' or 3' splice sites (alternative 5' splice site or alternative 3' splice site). In some other events, the introns fail to be removed by the spliceosome, hence included in the mRNA, which is called intron retention. More details about the different modes of AS can be found in Chapter 4.

Regulation of alternative splicing

Alternative splicing can be regulated by many elements (enhancer or silencer) located in both exons and introns. These elements can be bound by RNA binding proteins (RBPs), which act as activators or inhibitors of the splicing of the exons. For example, the SR (serine and arginine enriched) protein family consists of many well studied splicing factors (5,11). They can bind to exonic splicing enhancers (ESEs) and activate the splicing process. There are also negative regulatory elements in exons -- exonic splicing silencers (ESSs). The best characterized ones can bind to hnRNP proteins (heterogeneous nuclear ribonucleoproteins), which are splicing inhibitors (12).

There are many elements present in the intron region in addition to those in the exons. Some of them can even act from hundreds of nucleotides away from the splice sites. Like the regulatory elements in the exons, these elements are called intronic splicing enhancer (ISE) or intronic splicing silencers (ISS). These elements are usually conserved and can be identified by alignments of sequences between different species, but their functions are less characterized compared to exonic elements (5).

RNA splicing and sequence conservation

Many studies have shown a lifted conservation in the intron sequences adjacent to splice sites, due to the presence of ISE and ISS elements which regulate the splicing process. It is also noted that alternative spliced exons have higher conservation scores in both exons and flanking introns (13). The sequence conservation usually indicates functional alternative splicing events, and is correlated with tissue specificity (14).

Nonsense-mediated mRNA decay

Nonsense-mediated mRNA decay (NMD) is a pathway which can degrade mis-spliced or non-functional isoforms that contain pre-mature stop codons (PTCs) (15,16). In the process of mRNA splicing, a complex named exon-junction complex (EJC) is attached to the location of a removed intron. In the later translation process, when the ribosome goes through the mRNA to translate the sequence into protein, any proteins (including the EJC) in its path will be displaced. Upon its arrival at the stop codon, the ribosome will release factors with remaining EJCs (if any), which will trigger the degradation of the transcript. Generally, NMD triggering isoforms follow the 50-nt rule, which is, the stop codon should be at least 50 nt away from a downstream exon junction.

In some occasions, alternative splicing can introduce PTC to generate NMD triggering isoforms. For example, inclusion of an exon may shift the coding frame of a transcript and generate a PTC before 50 nt of an exon junction. This regulation is termed regulated unproductive splicing and translation (RUST), which is a means of indirect regulation of transcription by splicing factors (17). This aspect of regulation will be further discussed in Chapter 5.

Traditional techniques and methods for genome-wide alternative splicing study

Global insights into AS were initially achieved largely from analysis of expressed sequence tag (EST) data (18). Alternative splicing events can be identified by comparing different EST sequences. Generally, EST data are low coverage, expensive and have limited capability for quantifying exon-inclusion level, especially in specific conditions,

such as in different tissues. Other tag based methods were developed to overcome the limitations: including SAGE (serial analysis of gene expression), CAGE (cap analysis of gene expression) and MPSS (massively parallel signature sequencing). These techniques are high-throughput, however, most of them are based on Sanger sequencing technology and thus still expensive and the short tags sequenced cannot be uniquely mapped to the genome (19).

Another category of traditional methods is array based method, such as exon-junction arrays (20,21) or exon arrays (22). These methods take advantages of known gene structures and AS events observed in ESTs and other sequenced transcript data, and were designed to target the exon junctions or exon bodies. However, microarrays are largely restricted to studies of annotated AS events, and their signal-to-noise ratio is also limited by issues such as cross-hybridization. In addition, the dynamic detection range is limited due to both background and saturation of signals, and the normalization between samples can be difficult and need complicated methods.

Next generation sequencing technology and mRNA-Seq

Deep mRNA sequencing (mRNA-Seq) technology is a recently developed approach which uses the next generation sequencing to profile transcriptome at an unprecedented depth. Generally, the mRNAs are reverse transcribed into a library of cDNA fragments. Adaptors are attached to the fragments on one or both ends. Then the cDNA library is sequenced on a NGS platform. And millions of short reads are obtained from either one end (single end reads) or both ends (paired end reads) of the fragments.

The most important feature of mRNA-Seq technique is its ability to obtain ultra-high coverage of the trascriptome. Taking Hi-Seq 2000 from Illumina as an example, up to 55 Gb of nucleotides can be sequenced in a single day with the paired end 100 basepair (bp) run. This makes mRNA-Seq an ideal technique to profile transcriptome, especially to discover those lowly expressed isoforms which are relatively harder to detect with traditional methods. Of note, unlike microarray based methods, mRNA-Seq is able to detect novel gene structures which are not annotated in reference, providing opportunities to discover novel exon junctions and isoforms.

Another advantage of mRNA-Seq is the low background noise compared to microarrays. The dynamic detection range is much broader in mRNA-Seq since there is no upper limit of the quantification and it is able to observe either extremely abundant or lowly expressed transcripts (19).

There are also challenges brought by some of these features, for example, the RNA-Seq reads are relatively shorter than the tags in Sanger sequencing, and makes it hard to identify junctions from the reads. These challenges will be discussed in the following section.

Current methods to analyze mRNA-Seq data

Alignment of the exon junction reads

The first step to analyze mRNA-Seq data is the mapping of the reads back to the genome to locate exon body reads or exon junction reads. On one hand, shorter reads increase the possibility of getting multiple hits on the genome, and also makes it harder to identify exon junctions due to the lacking of sequence overlaps in the exons. On the other hand, the large amount of data requires optimizations in the algorithm to make sure that the alignments can be done in a reasonable time frame.

Exon body reads are relatively easier to map, because there is presumably no large gaps (introns) in the reads. There are many programs developed for this purpose. The first generation of the aligners (Eland, Maq, SOAP, RMAP, ZOOM, and SHRiMP)(23-26) use hashing table to store the reads or the reference, which is computational expensive. Later, Burrows-Wheeler transform (BWT) and the FM index(27) were introduced into NGS data alignment by BWA, Bowtie and SOAP2(28-30). The BWT algorithm enables efficient backward search with small memory footprint when there are few mismatches between the read and the reference. However, all these algorithms are designed for continuous alignment without big gaps, hence are not suitable for splicing study.

Compared to exon body reads, exon junction reads are more difficult to identify, however, they are direct evidence of splice sites and are important for AS study. There

were many algorithms developed in the recent years to specifically align spliced reads to exon junctions, these include earlier methods which constructed databases of known junction sequences, e.g. ERANGE (31), and later methods which can identify novel exon junctions using different heuristics, such as TopHat, MapSplice, SpliceMap, SOAPsplice, PASSION, HMMsplicer and GSNAP(32-38). To locate exon junction reads, a common approach used by most of the programs here is the seed-and-extend method. Using this method, the programs first split the read into two or more segments (seeds) and align them onto the genome separately. Then the junction searching is performed between two aligned seeds (double-anchor search) or around a single seed alignment (single-anchor search) . However, most of these methods require external mappers (like Bowtie) and limit the use of short seeds, because they need to control the temporary files within an acceptable size. This limits their sensitivity on exon junction detection. To overcome this drawback, we designed a novel algorithm and program named OLego(39) for fast and sensitive mapping of mRNA-Seq reads, which will be further described in Chapter 2.

Reconstruction of gene structures

Although a number of transcriptomes from many species (like human and mouse) have been extensively studied in the past decades, which can be used as reference for quantification, there are still many transcripts that are not annotated in databases. Hence, a major task after the mapping of mRNA-Seq reads is the reconstruction of the gene structures. In this stage, the reads are assembled back to transcriptome to find the different mRNA isoforms. There are many reasons making the transcriptome reconstruction a difficult problem(40). First of all, the reads are relatively short, making it a challenge to allocate the reads to an isoform, since many isoforms may share the same region. Secondly, some transcripts are very low abundant, giving little evidence of their structures. Thirdly, there are noises coming from the pre-mRNAs, which might fall into both introns and exons, further increasing the complexity of the problem.

There are two categories of methods developed to reconstruct the transcriptome from the reads. One category is the genome-dependent method, like Cufflinks and Scripture (41,42), both of which use the genome sequence as a reference to do the reconstruction. And the other category is genome-independent method, e.g. Trans-

ABySS(43), which assembles transcriptome without requiring reference genome. Obviously, genome-dependent methods are preferred when the reference genome is available, e.g. human and mouse, since the genome-independent methods are more computational expensive and less sensitive in this case. But this also depends on the biological questions focused.

Generally, these reconstruction methods aim to find a balance between accuracy and sensitivity. On one hand, they want to recover as many as isoforms to make the prediction as close as possible to the truth, while on the other hand, they try to explain the reads with as few transcripts as possible. For example, Cufflinks and Scripture, both of which are popular algorithms designed to build assembly graphs based on the alignments, while they differ in how to report the isoforms: Cufflinks reports the minimum isoform sets to explain the reads (maximum precision), and Scripture reports all possible isoforms based on the reads (maximum sensitivity). There are also other programs developed in the recent years, such as IsoLasso, SLIDE and NSMAP (44-46), which use LASSO-based (LASSO: Least Absolute Shrinkage and Selection Operator) approaches to find major isoforms of a gene.

For AS study, it is not always necessary to reconstruct the whole transcripts to identify the splice events. In most occasions, the local information of the reads around the exons is enough for recovering the AS events involving the exons. Typically, due to the complexity of the transcriptome, the local structures are shared by multiple transcripts. To reconstruct the transcripts, people need to further connect these local structures to recover the whole structure. But for splicing study, we can ignore the distant information and identify AS events solely with local structures. This will be the major topic of Chapter 3.

Expression quantification from mRNA-Seq

There are two levels of expression quantification from mRNA-Seq. The first level is the gene expression, which does not consider the different isoforms expressed at each gene locus but rather the whole gene. The first and intuitive method is the RPKM (reads per kilobase of transcript per million mapped reads) approach (31), which normalizes the

read counts at a certain locus by both gene length and the total number of mapped reads in the experiment. The second level of the quantification is the transcript expression, in which the reads are assigned to different isoforms of a gene for more detailed expression measurement. This type of quantification is more useful for researchers working on the splicing level.

The simplest way to quantify isoforms from mRNA-Seq reads is the count based method. For example, ALEXA-seq(47) is an approach only using uniquely mapped reads which can be located into a single isoform, this largely limits the reads used in the estimation and also misses transcripts without uniquely mappable regions. More sophisticated methods solve the problem in more decent ways by modeling different properties of RNA-Seq data, such as insert size distribution for paired-end data, non-uniformity of the reads along the transcripts, and sequence bias at both ends of the reads. The statistical models then allocate the reads into different isoforms to estimate the abundance of each isoform, e.g. Cufflinks(41) and MISO (mixture of isoforms)(48). Comprehensive discussion and comparison of the models can be found in a review written by Lior Pachter (49).

For AS study, a natural next step would be the measurement of splicing from the expression profile of the transcriptome. For example, to estimate the inclusion ratio of a certain exon, we can use the expression level of the isoforms containing this exon and the expression level of those without this exon to compute the ratio. However, there is a conflict between sensitivity and accuracy, since sometimes some lowly expressed isoforms are either ignored or quantified with low accuracy. To measure AS in a more accurate and sensitive way, we developed SpliceTrap(10), which works in local regions to estimate exon inclusion ratios. This will be discussed in Chapter 4 in more details.

Organization of the dissertation

The rest of the dissertation is organized in the order of the steps in data analysis:

Chapter 2: OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. This chapter describes the first step to analyze RNA-Seq data to find the locations of the reads, especially those reads crossing exon junctions, which are essential to RNA splicing study.

Chapter 3: Identification of exons and splicing events from alignment results. This chapter briefly describes our scripts to extract alternative splicing events directly from the mapping results.

Chapter 4: SpliceTrap: a method to quantify alternative splicing under single cellular conditions. This chapter shows a method to quantify alternative splicing based on the splicing events discovered from the data using a Bayesian model.

Chapter 5: Systematically discovery of conserved alternative spliced exons in the mammalian genome, which demonstrates an application of the whole pipeline on a real data, together with a set of post analysis.

Author contributions

This dissertation describes my major work during my Ph.D. study, while there are other people who contributed significantly to the projects. Martin Akerman is co-first author of the SpliceTrap paper. He worked closely with me on improving the program, carrying out the simulation study, presenting the data and writing the paper. Shuying Sun generated the RNA-Seq data used by SpliceTrap paper. The OLego project was first initiated by Chaolin Zhang, who also advised me in the whole project. Olga Anczuków helped to carry out the RT-PCR validations of the micro-exons identified in the mouse retina RNA-Seq data. The BodyMap project (Chapter 5) was based on Chaolin Zhang's previous work about conservation of alternative splicing and is under his supervision.

Chapter 2 :

OLego: Fast and Sensitive Mapping of Spliced mRNA-Seq Reads Using Small Seeds

Abstract

A crucial step in analyzing mRNA-Seq data is to accurately and efficiently map hundreds of millions of reads to the reference genome and exon junctions. Here we present OLego, an algorithm specifically designed for *de novo* mapping of spliced mRNA-Seq reads. OLego adopts a multiple-seed-and-extend scheme, and does not rely on a separate external aligner. It achieves high sensitivity of junction detection by strategic searches with very small seeds (~14 nt for mammalian genomes). To improve accuracy and resolve ambiguous mapping at junctions, OLego uses a built-in statistical model to score exon junctions by splice-site strength and intron size. Burrows-Wheeler transform (BWT) is used in multiple steps of the algorithm to efficiently map seeds, locate junctions, and identify very small exons. OLego is implemented in C++ with fully multi-threaded execution, and allows fast processing of large-scale data. We systematically evaluated the performance of OLego in comparison with published tools using both simulated and real data. OLego demonstrated better sensitivity, higher or comparable accuracy, and substantially improved speed. OLego also identified hundreds of novel micro-exons (< 30 nt) in the mouse transcriptome, many of which are phylogenetically conserved and can be validated experimentally *in vivo*. OLego is freely available at http://zhanglab.c2b2.columbia.edu/index.php/OLego.

Introduction

In eukaryotes, alternative splicing (AS) is critical for amplifying genomic complexity by generating multiple mRNA isoforms from a single gene (5,50). More than 90% of human multi-exon genes express transcripts that potentially undergo AS (8,9). Besides the extent of AS, decades of research have revealed the key roles of this process in post-transcriptional gene-expression regulation, and how its disruption can cause various genetic diseases (51,52)

Global insights into AS were initially achieved largely from analysis of expressed sequence tag (EST) data, which provide a means of cataloguing AS events at the genome-wide scale (18). In general, EST data have low coverage and limited capability for quantifying exon-inclusion level, especially in specific conditions, such as in different tissues. This issue was later addressed by splicing-sensitive microarrays, such as exon-junction arrays (20,21) or exon arrays (22), which were designed based on gene structures and AS events observed in ESTs and other sequenced transcript data. However, microarrays are largely restricted to studies of annotated AS events, and their signal-to-noise ratio is also limited by issues such as cross-hybridization. Recently, ultra-high throughput mRNA sequencing (mRNA-Seq) provided a powerful alternative to profile the transcriptome at unprecedented depth and resolution, with the advantages of being highly quantitative, sensitive, and able to discover novel splice junctions and exons (19).

A key step in analyzing mRNA-Seq data is to map hundreds of millions of reads, currently of size 50-150 nucleotides (nt), back to the reference genome, and to detect known or novel splice junctions. Various algorithms have been developed in the past few years for this purpose, with specific consideration to mapping speed and to short read lengths (32-35,38,53,54). The early versions of TopHat (32) first align all exon-body reads to the genome using an external aligner, Bowtie (29), and all aligned reads are clustered and counted to locate potential exons based on read coverage (exon islands). Potential splice sites are then searched locally, and nearby exons are paired *in silico* to generate a database of candidate exon junctions, followed by alignment of unmapped reads in the first stage against this junction database. This procedure is relatively fast and reliable, because exon identification prior to junction search largely limits the search

space, despite the caveat that junctions spanning exons at low levels might be missed. To overcome this limitation, several other programs turned to more exhaustive searches by using double- or multiple-seed-and-extend approaches to find exon junctions *de novo*. For example, SpliceMap (35) splits each read of ~50 nt in the middle and maps each part (seed) to the genome separately, again relying on an external aligner for genomic mapping, and then it extends the alignments to find junctions. To handle longer reads that can span multiple junctions—obtained with more recent technologies—MapSplice (33) and later versions of TopHat (32) segment each read into multiple seeds to detect splice junctions.

Although different heuristics are used in each algorithm, an important limitation shared by these tools is their use of relatively long seeds (~25 nt). This is due in part to their dependence on an external aligner for seed mapping, whose output is then parsed to detect exon junctions. As a consequence, the number of hits for each seed has to be small, which constrains the choice of seed size and limits the resolution in locating potential exon positions. This constraint increases the chance that one or more seeds will fail to align, because they span exon junctions, reducing the sensitivity of junction detection. This issue becomes more severe for reads spanning small exons, which are frequently alternatively spliced and regulated to have variable inclusion levels in specific conditions.

As sequencing technologies keep evolving, the throughput and read length are increasing very rapidly, which imposes even greater challenges for mRNA-Seq data processing. For example, a single sequencing lane from the Illumina HiSeq 2000 can currently produce over 200 million paired-end reads, with read lengths up to 150 nt. Therefore, mapping speed, without sacrificing accuracy, becomes more critical. In addition, longer reads tend to span more exon junctions and have more complex structures, especially when they cover small exons or exons expressed at a low level. Here we address these challenges and present a new program named OLego, which is designed for very fast, *de novo* mapping of spliced mRNA-Seq reads with both high specificity and sensitivity.

Materials and methods

Overview of mRNA-Seq read mapping

Analysis of mRNA-Seq data typically starts from mapping a set of *N* relatively short reads of length *L* nt to the reference genome. For higher eukaryotes, and mammals in particular, the vast majority of genes consist of multiple exons and introns. Therefore, a read can be mapped continuously to a single exon (exonic alignment), or to multiple exons that span one or more exon junctions (junction alignment). Due to sequencing errors or polymorphisms in the sequenced sample, compared to the reference genome, a read alignment has to tolerate a certain number of substitutions or small insertions and deletions (indels)—collectively denoted as mismatches here—as measured by an editing distance of *M* nt between the query reads and the target reference genome sequences.

OLego finds junction alignments using a multiple-seed-and-extend approach, which is also used by several other programs, such as MapSplice (33), but with several distinct and important features (Figure 1). In essence, each read is processed independently in a series of steps, without relying on an external aligner. Reads can therefore be processed in parallel when multiple threading is enabled. OLego performs more exhaustive and yet efficient searches using very small seeds (12-14 nt; 14 nt for this study), whose hits are clustered, ranked, and refined to find the best alignment. This greatly improves the sensitivity for *de novo* discovery of splice junctions and small exons. In addition, particular attention is paid when a small unaligned segment of a read is flanked by aligned regions on both ends in the presence of large genomic gaps, typically due to the presence of a very small exon (<30 nt) or micro-exon (55). To ensure the efficiency of this exhaustive procedure in terms of both time and memory usage, Burrows-Wheeler transform (BWT) and FM-index (56) are used in multiple steps to map seeds and discover junctions and small exons with a small memory footprint (< 4Gb in general for mammalian genomes). More details of the algorithm are described below.

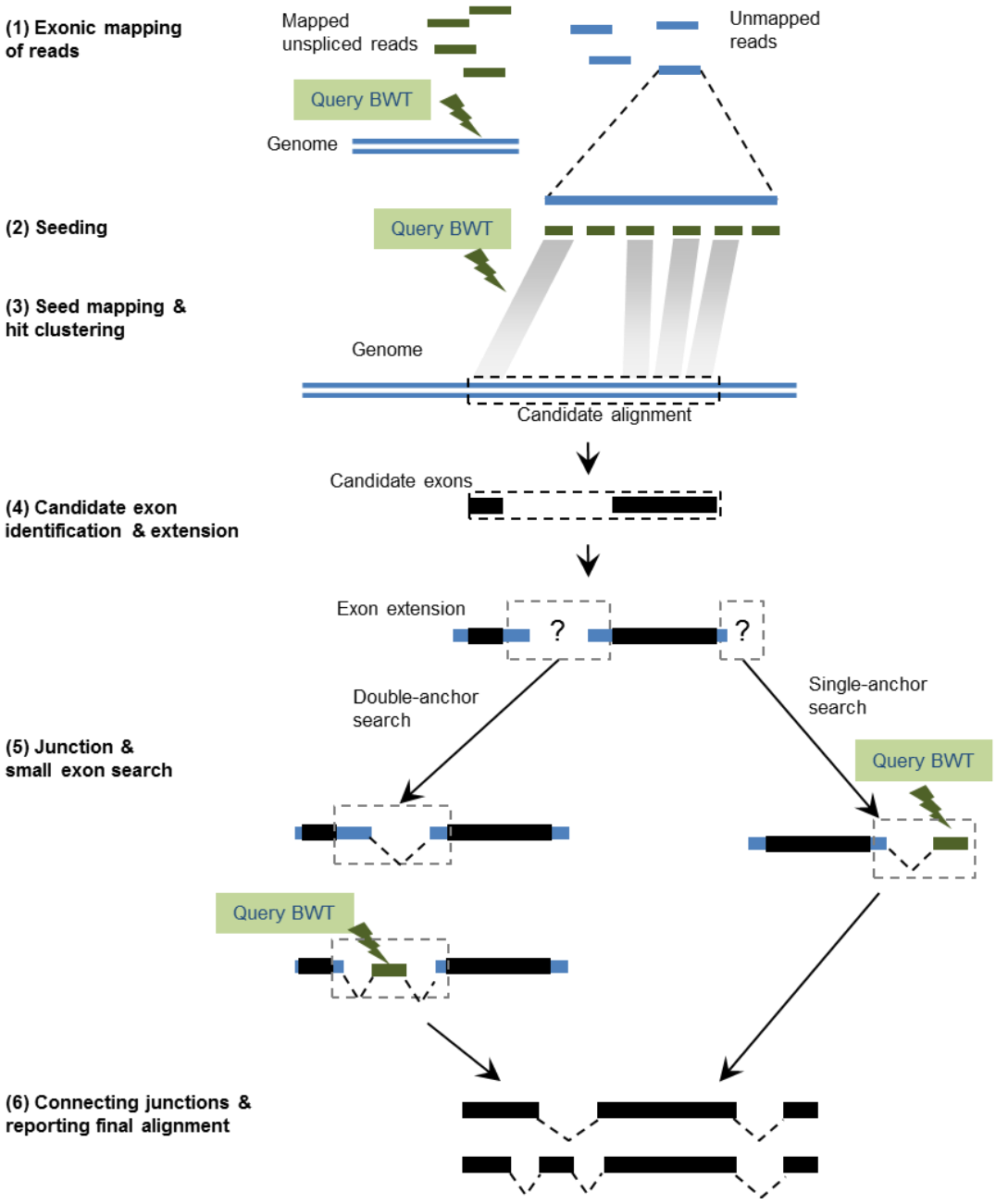


Figure 1. Overview of OLego. Each read is processed independently by OLego. (1) Continuous mapping to the genome or exonic alignment is attempted first. If no hits are found within the allowed number of mismatches, junction alignment is searched through steps starting from (2) seeding (3) seed mapping and hit clustering into candidate alignments, and (4) candidate-exon identification and extension. (5) Junctions are then searched between two consecutive candidate exons and at the end of the read, and small exons are searched when necessary. (6) Finally, exons and junctions are connected and ranked to identify the optimal alignment for the whole read.

The workflow of OLego

1. Exonic mapping of reads. For each read, continuous mapping to the genome with BWT and FM-index is first attempted, using essentially the same approach as in BWA (28), with minor modifications. At most M' nt (currently $M'=\min\{2,M\}$, where M is the number of mismatches allowed for the whole read) mismatches are allowed in this step. If an exonic alignment is found, the read will be reported as an exon-body read, and the algorithm turns to the next read. Otherwise, it will be processed in the following steps to search for a junction alignment. Note that a smaller number of mismatches are allowed here to avoid promiscuous exonic alignment with mismatches near the end of a read, when it actually spans an exon junction with a small anchor at the end. Exonic alignment with $>M'$ mismatches (but $\leq M$ mismatches) will be recovered later (step 4 below).

2. Seeding. Each unmapped read subject to junction search is segmented into multiple seeds of a specified size w. Spaces are allowed between seeds if the read length is not a multiple of the seed size, such that the read can be evenly covered by seeds. Since the boundaries of seeds relative to exon junctions are random, an exon of size $\geq 2w$ is guaranteed to have at least one seed inside the exon, assuming a sufficient sequencing depth. The default seed size w in OLego is 14 nt, considering the balance between sensitivity and speed to deal with mammalian-sized genomes. The use of a smaller seed size in OLego greatly increases the chance of finding hits of one or more seeds in each exon, especially for small exons $\leq$50 nt.

3. Seed mapping and hit clustering. Each seed is mapped independently to the genome by querying BWT and FM-index, allowing $\leq m$ mismatches (default: $m=0$). Due to the small seed size, each seed is expected to have a substantial number of hits. For example, at a seed size of 14 nt, the average number of hits for each seed is estimated to be $W=11$ for a mammalian genome ($3\times10^9/4^{14}$), although this number varies for different seeds. If a seed has an exceedingly large number of hits ($W>1000$), it is considered as repetitive and all its hits are discarded; otherwise, we keep all W hits of a seed, and recover their original genomic coordinates from the BWT index. The hits of all kept seeds are then clustered head-to-tail to locate potential alignments of the complete read according to their genomic coordinates, so that the distances between any two neighbor

hits in each potential alignment are less than twice the specified maximum intron size *I* (default: *I*=500,000 nt). We require *2I* in the clustering of hits, because there might be a missing internal exon between two neighbor hits (see below). Each potential alignment is scored and ranked according to an "E-value" estimated from the number of aligned seeds and their uniqueness $E = G\prod_i (o_i/G)$, where o_i is the total occurrence in the whole genome for seed *i* in the potential alignment, and *G* is the size of the genome. Only the top 100 potential alignments with $E<10$ are examined further.

To maximize the speed, we do not allow mismatches in seed mapping by default, given the small seed size and low sequencing errors that minimize the chance of failure in seed mapping. In addition, even if no hits are found or kept for some seeds, due to sequencing errors, polymorphisms, or their repetitive nature, such parts can still be recovered in the following hole-filling and candidate-exon-extension step. Each potential alignment is treated separately in the following steps.

4. Candidate-exon identification and extension. In each potential alignment, the hits are further grouped into individual candidate exons, using more stringent criteria. This is done using the diagonal coordinates of the hits, which are calculated by subtracting the start coordinates of the corresponding seeds in the query read from the genomic start coordinates of the hits (57). The hits whose diagonal coordinates are within *M'* nt differences are considered to be in the same candidate exon, which tolerates potential indels in mRNA-Seq reads. After this step, holes between hits within each candidate exon are filled in by realigning the orthologous sequences in the query read and the reference genome using banded dynamic programming, which allows substitutions and small indels. In addition, each candidate exon is also extended on both ends by allowing $\leq M'$ mismatches to find potential exon boundaries. If a candidate exon already covers the whole read with $\leq M$ mismatches at this point, a candidate exonic alignment is recorded.

5. Junction and small-exon search. There are two types of junction searches: double-anchor and single-anchor. Double-anchor search is performed between each pair of neighboring candidate exons. Candidate splice sites are searched locally around the exon boundaries (default: ±6 nt). At the same time, the match of sequences between the reference genome and the query read around exon boundaries are examined. If

nucleotides near exon boundaries are aligned properly ($\leq M'$ mismatches), a candidate exon junction is recorded. Otherwise, if a gap remains in the query read after local search of exon boundaries according to the candidate splice sites, this typically suggests a missing internal exon without any hits of seed sequences in the exon, as discussed above. In this case, further searching for the missing internal exon is carried out. The sequence in the gap region of the read, flanked by the dinucleotides (AG/GT) of the two splice sites, is queried against the reference genome using BWT and FM-index to find the missing internal exon, requiring a minimum exon size (default: 9 nt) and proper intron size (default: 20 nt ~ 500,000 nt). This gives a chance of approximately 5.8×10^{-5} ($2 \times 500{,}000/4^{9+8}$, 8 nt are dinucleotides for four splice sites) to find a random match.

Single-anchor search is performed at the ends of the first and last candidate exons if they do not reach the boundaries of the read. Candidate 5 ´or 3 ´splice sites are searched locally (default: ±6 nt) near the exon boundary, and the unaligned part of the read after this local adjustment, flanked by the 3 ´or 5 ´splice site dinucleotide, is searched against the reference genome with BWT and FM-index. The size of the match at the end has to be larger than the minimum size (default: a=8 nt), and the intron size is restricted in the proper range as well. This gives approximately a chance of 0.03 ($500{,}000/4^{8+4}$, 4 nt are splice site dinucleotides) to find a random match.

6. Connecting junctions and reporting the final alignment. All candidate junctions are connected along the read to find the optimal path that represents the complete alignment of the whole read. If multiple candidate alignments can be found within the desired number of mismatches, all candidate alignments are first ranked according to the number of mismatches. If the top two or more alignments have the same number of mismatches, they are further ranked to resolve ambiguity by an additional criterion that takes into consideration splice-site strength and intron size. This criterion is also used to filter out potential false positives in *de novo* junction search (details are given below).

A regression model to score exon junctions

Splice sites show extended consensus sequences beyond the strictly required GT/AG dinucleotides (for canonical splice sites), which are crucial for accurate and efficient exon

recognition by the splicing machinery (58,59). These motifs have been used previously for bioinformatic splice-site prediction in several methods, such as GeneSplicer (60) and SplicePort (61). In addition, intron size also affects the efficiency of splicing, and shows a distinct distribution in the mammalian genome (62).

When multiple alignments with the same number of mismatches exist, they are further ranked to prioritize the most reliable alignments according to the strength of exon junctions, using a regression model that combines splice-site motif score and intron size. To this end, we collected true splice sites from annotated gene models. For example, for mouse data, NCBI37/mm9 Ensembl (63) gene annotations were downloaded from the UCSC genome browser (64); all the splice-site pairs (242,141 pairs) were then retrieved as a true-positive training dataset. We also randomly selected the same number of pairs of GT/AG sites separated by 20-500,000 nt from the mouse genome to generate the training dataset of false splice sites.

The splice-site score for each exon junction is calculated using ±15 nt sequences around the 5´and 3´splice sites, respectively (65). Therefore, for each pair of splice sites corresponding to an exon junction, 60 nt are taken into account. We define the splice-site score *S* of an exon junction as:

$$S = \sum_i s_{i,B_i} = \sum_i \log\left(p_{i,B_i} / p_{0,B_i}\right) \qquad (1)$$

where B_i is the nucleotide (A, C, G or T) at position i, p_{i,B_i} is the probability of observing B_i at position i of the 60-nt splice-site motif derived from the true dataset, and p_{0,B_i} is the probability of observing B_i in the background intronic sequences. Junction splice-site scores are calculated for all the entries in both true and false training datasets. Meanwhile, the corresponding intron sizes are recorded.

Splice-site score and intron size are combined by a logistic function:

$$f(z) = \frac{e^z}{e^z + 1} \qquad (2)$$

and

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (3)$$

Here x_1 and x_2 are splice-site score and intron size, respectively, and the coefficients are determined by fitting the true and false training datasets. We have

provided the regression models for mouse and human in the package, and the parameters of these models, including the coefficients and their statistical significance, are summarized in Supplementary Table 1). Scripts are also included to allow users to generate their own models for other species and gene annotations.

For every candidate junction identified by single- or double-anchor *de novo* junction search, we calculate the logistic probability with Equation 2. At the "junction connection" stage, logistic probabilities of all junctions in a candidate alignment are averaged for final ranking. Those *de novo* alignments with low logistic probabilities (default: ≤ 0.5) are regarded as low confidence and filtered out.

Practical considerations in implementation

OLego can either perform *de novo* junction searches or work with a database of annotated splice junctions. For *de novo* junction search, we currently require the canonical GT/AG splice sites, because they account for ~99% of all known introns in mammals (66). To further reduce false-positive detection of splice sites, we also require an average logistic probability of >0.5 for each alignment. When OLego is provided with a database of annotated splice junctions, several special considerations are given for alignment to known splice sites, because they define a much smaller search space. Specifically, we allow non-canonical splice sites, a less stringent threshold on the minimum anchor size (5 nt vs. 8 nt for single-anchor search), and no constraint on intron size.

OLego takes FASTA or FASTQ files as input, and outputs alignments in SAM format (67). The junctions from the best alignments are collected and reported in BED format. It loads mRNA-Seq reads in batches, and in each batch the reads are assigned randomly to different threads, when multiple threading is enabled. Therefore, OLego supports multiple threading in the whole alignment workflow. This is distinct from many available tools, for which multiple threading is only supported at the stages when an external aligner is involved. For paired-end mRNA-Seq data, each end is first mapped independently, and the results for both ends are then combined according to their distance and orientations on the reference genome, to help resolve possible ambiguity in alignments of single-end reads. Different types of mRNA-Seq libraries with or without strand information can be handled properly.

OLego is an open source code project. It is released under GPLv3 and is freely available online at http://zhanglab.c2b2.columbia.edu/index.php/OLego. It was implemented in C++ and relied heavily on the source code library of BWA (version 0.5.9rc1) (28).

Evaluation on simulated datasets

We generated simulated mRNA-Seq reads using the program BEERS from the RUM package (53). For mouse (mm9), BEERS uses gene models derived from 11 annotation tracks (AceView, Ensembl, Geneid, Genscan, NSCAN, Other RefSeq, RefSeq, SGP, Transcriptome, UCSC, and Vega) in the UCSC genome browser to avoid bias towards or against any particular set of gene annotations. It is also capable of simulating polymorphisms and random sequencing errors (at a default rate of 0.5%) with positional biases (e.g., higher error rate towards the end of reads) that mimic real mRNA-Seq data produced by the Illumina platform. We carried out two sets of simulations with default parameters, each consisting of 10 million paired-end reads, but with different read lengths (100 and 150 nt, respectively); three replicates were generated for each set.

We compared OLego (v1.0.0) with three other published programs: TopHat (version 1.4.0), MapSplice (version 1.15.2) and PASSion (version 1.2.1). TopHat and MapSplice use seed-and-extend approaches, as described above. Alternatively, PASSion (34) uses a different strategy, called pattern growth, which does not require segmentation of reads, to find exon junction reads in paired-end mRNA-Seq data. For all these programs, default parameters were used for mapping, except that the size of introns was restricted in the range of 20-500,000 nt for OLego, TopHat and MapSplice, and 20-409,600 nt for PASSion due to its discrete choices for the maximum intron size. In addition, up to 4 mismatches were allowed by OLego (-M 4). In this setting, OLego searches with a seed size of 14 nt (-w 14), allowing no mismatches in the seed; *de novo* single-anchor junction search is enabled and a minimum anchor size of 8 nt is required (-a 8). For MapSplice, the configuration file paired.cfg included in the package was used to maximize the sensitivity (see Discussion). Both MapSplice and Tophat used a seed size of 25 nt and minimum anchor size of 8 nt, and they tolerated 1 and 2 mismatches in

the seed, respectively. The reads were mapped onto the reference mouse genome (mm9) without any exon junction annotations provided. Up to 16 Intel Xeon CPU cores (2.0 GHz) on a Linux server were used for mapping. The BED format junction output files from these programs were used to evaluate discovery of unique exon junctions, and the alignment outputs (in SAM or BAM format) were used to evaluate the accuracy of junction alignment and small-exon discovery.

Evaluation on real mRNA-Seq data

We downloaded mRNA-Seq data (accession: SRX088978) used in a previous study (53) from the NCBI Sequence Read Archive (SRA) (68). This mouse retina mRNA-Seq library was originally prepared with a 350 nt (±25 nt) average insert size, and sequenced on an Illumina Genome Analyzer IIx, with 120 nt paired-end reads (53). One lane of reads (~26 million reads) was extracted and used in our study. Parameters used in OLego for read alignment were the same as those used in simulation, as described above.

RT-PCR validation of novel micro-exons

Retinal tissues from three two-month-old female C57BL/6J mice were purchased from The Jackson Laboratory. Total RNA was extracted using Trizol (Invitrogen), followed by DNAse I digestion (Promega), phenol-chloroform extraction, and ethanol precipitation. 1 μg of RNA was reverse-transcribed with Improm-II reverse transcriptase (Promega) and oligo dT primers.

Radioactive touchdown PCR with [α-32P]-dCTP and Taq Gold polymerase (Invitrogen) was used to amplify endogenous transcripts with primers described in Supplementary Table 4 and Figure 7A. PCR products were separated by 8% native PAGE, visualized by autoradiography, and quantified on a phosphorimager (Fuji Image Reader FLA-5100) using Multi Gauge software Version 2.3. The inclusion ratio of each exon was then calculated by normalizing the signal intensity of the inclusion isoform to the total intensity of both isoforms, and expressed as a percentage.

Results

Exon junction discovery in simulated datasets

We first evaluated the performance of OLego for *de novo* exon junction discovery using two sets of simulated mRNA-Seq data. We also carried out a comparison of OLego with two other widely used seed-and-extend programs, TopHat (32) and MapSplice (33), and a recently published program, PASSion, which is based on a pattern-growth algorithm to search exon junctions in paired-end data (34). All of the compared programs were previously benchmarked and demonstrated good performance (34,53). In each simulation set, we generated 10 million 100-nt or 150-nt paired-end reads, which were aligned by the four programs. Unique exon junctions reported by each program were used to estimate the positive predictive value (PPV) as a measure of accuracy, and false negative rate (FNR) as a measure of sensitivity, and this process was repeated in three replicates to average the results (Table 1A). In these tests, a slightly higher PPV was achieved by Tophat and OLego (97.7-98.4%), compared to PASSion (96.3% for both 100-nt and 150-nt reads) and MapSplice (95.1% for 100-nt reads; 97.1% for 150-nt reads). In terms of sensitivity, OLego discovered substantially more true junctions than the other programs. OLego's FNR (8.2% for 100-nt reads and 6.8% for 150-nt reads) almost halved the FNRs of TopHat (15.4% for 100-nt reads and 12.8% for 150-nt reads) and PASSion (14.8% for 100-nt reads and 15.5% for 150-nt reads), whereas MapSplice had an intermediate FNR (10.3% for 100-nt reads and 9% for 150-nt reads). Therefore, OLego achieved both high sensitivity and specificity, suggesting the benefit of more exhaustive searches using very small seeds, combined with quantitative modeling of exon-junction strength and alignment quality. As expected, all seed-and-extend based tools achieved better sensitivity when the read size increased; interestingly, PASSion's sensitivity decreased slightly with longer reads.

Table 1. Exon junction discovery by OLego, MapSplice, TopHat and PASSion on simulated data.

A. The number of exon junctions identified by each program.

	100-nt reads (178,449 junctions)				150-nt reads (189,106 junctions)			
	OLego	**MapSplice**	**TopHat**	**PASSion**	**OLego**	**MapSplice**	**TopHat**	**PASSion**
Found junctions total	166,954	168,219	153,446	157,967	180,410	177,142	167,707	165,871
Found true junctions	163,740	159,984	151,013	152,094	176,172	172,052	164,847	159,773
Missed true junctions	14,708	18,464	27,436	26,355	12,934	17,054	24,259	29,333
PPV	0.981	0.951	0.984	0.963	0.977	0.971	0.983	0.963
FNR	0.082	0.103	0.154	0.148	0.068	0.090	0.128	0.155

B. The observed (upper diagonal; shaded) and expected (lower diagonal) overlap of is covered true exon junctions between each pair of programs.

	100-nt reads (178,449 junctions)				150-nt reads (189,106 junctions)			
	OLego	**MapSplice**	**TopHat**	**PASSion**	**OLego**	**MapSplice**	**TopHat**	**PASSion**
OLego	163,740	156,654	147,731	147,266	176,172	169,122	161,421	155,259
MapSplice	146,798	159,984	148,671	146,290	160,285	172,052	162,297	154,520
TopHat	138,566	135,387	151,013	139,992	153,573	149,981	164,847	150,239
PASSion	139,558	136,357	128,710	152,094	148,845	145,364	139,277	159,773

PPV: positive predictive value or precision; FNR: false negative rate.

We then assessed the extent of overlap among the four programs with regard to the true junctions they identified. In all pairwise comparisons, the number of common junctions identified by the programs was higher than expected by chance (Table 1B, upper diagonal vs. lower diagonal). For example, in 100-nt reads, OLego identified most true junctions found by MapSplice (97.9% or 156,654/159,984), TopHat (97.8% or 147,731/151,013), and PASSion (96.8% or 147,266/152,094) whereas only 91.1%~91.8% (146,798/159,984, 138,566/151,013, and 138,558/152,094, respectively) were expected ($P<2.2\times10^{-16}$, Fisher's exact test). The striking statistical significance of the overlap suggests that some junction reads are easier to align, whereas others are more difficult for all four programs. This observation can be interpreted in several ways, including

multiple hits of read sequences in the transcriptome (e.g., introduced by paralogous genes), ambiguity of sequence alignment at exon junctions, short anchors on either side of some exon junctions, or complications introduced by simulated sequencing errors in some mRNA-Seq reads.

We also compared how read coverage affected each program in sensitivity of exon-junction detection. For this purpose, we binned the simulated junctions according to their ground-truth read coverage, and for each program, we estimated the sensitivity of junction discovery separately for each bin (Figure 2). As expected, all programs had higher sensitivity when the coverage increased. OLego achieved higher or comparable sensitivity in all bins, relative to the other three programs. For example, for 100-nt reads, OLego had a sensitivity of 95.7% for junctions supported by >4 reads, which was comparable to MapSplice (95.8%), despite more specific junction identifications by OLego. OLego performed best in all other bins, with sensitivity between 74.8% (for junctions supported by only 1 read) and 93.9% (for junctions supported by 4 reads). On the other hand, TopHat and PASSion had relatively lower sensitivity, as observed from all bins. Importantly, the advantage of OLego in sensitivity was particularly clear for exon junctions with low coverage, compared to the other three programs (63.7%, 48.3% and 54.1% for junctions supported by only 1 read for MapSplice, TopHat and PASSion, respectively; Figure 2), again suggesting the benefit of more exhaustive searches using very short seeds.

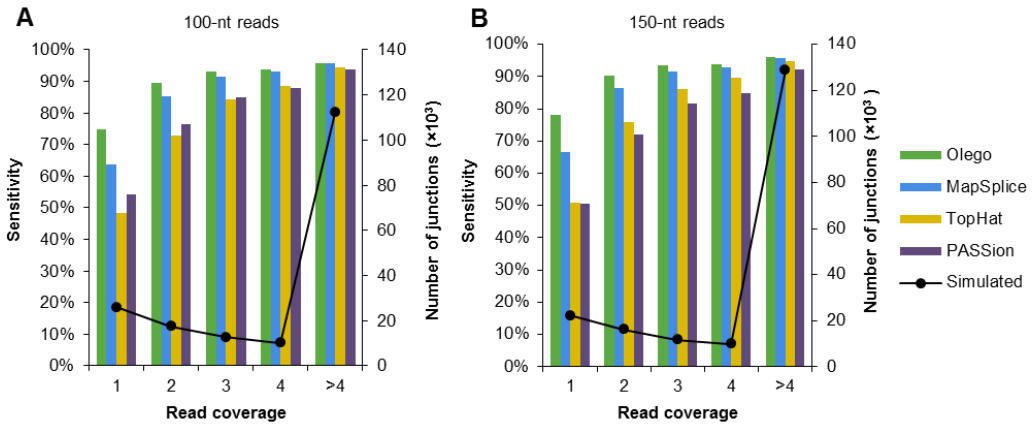


Figure 2. Sensitivity of junction detection at different coverages. (A) Tests on 10 million 2×100-nt simulated reads; (B) Tests on 10 million 2×150-nt simulated reads. For each panel, the simulated junctions were binned according to their coverage, from 1 read per junction to >4 reads per junction. The true numbers of junctions in the simulation are shown by lines with markers on the right axis, and the sensitivity of OLego, MapSplice, TopHat and PASSion are indicated by bars on the left axis.

Mapping speed

We next compared OLego, MapSplice, TopHat and PASSion in terms of mapping speed, because this becomes increasingly critical as the throughput of mRNA-Seq technologies increases dramatically. OLego supports multiple threading in the whole cycle of mapping individual reads, whereas the other three programs support multiple threading in a limited number of steps. Therefore, we first ran all programs with multiple threading enabled using 16 CPU cores (2.0 G Hz per core). It took OLego, TopHat, MapSplice and PASSion 0.8, 2, 5.5, and 6.8 hours, respectively, to align the 10 million 2×100-nt reads, and 1.4, 3.3, 10.5 and 9.3 hours, respectively, to align the 10 million 2×150-nt reads. OLego was faster than TopHat by more than 2-fold, whereas MapSplice and PASSion were substantially slower than OLego by about 7-fold. To further compare read-mapping speed and the benefit of multiple threading, we ran OLego and TopHat using different numbers of CPU cores (1, 4, 8 and 16) on both sets of simulated data (Figure 3). When a single CPU core was used, OLego and TopHat had similar mapping speeds, despite the fact that OLego performed more exhaustive searches using much smaller seeds. When more CPU cores were used, the mapping speed of OLego increased linearly as a function of the number of CPU cores. The mapping speed of TopHat also increased, but at a

slower rate. This is presumably because TopHat supports multiple threading only in steps that involve the external aligner Bowtie. When ≥8 CPU cores were used for alignment, OLego used half as much or even less time, compared to TopHat. These comparisons suggest that OLego not only achieved high sensitivity and specificity, but also substantially improved the mapping speed.

\

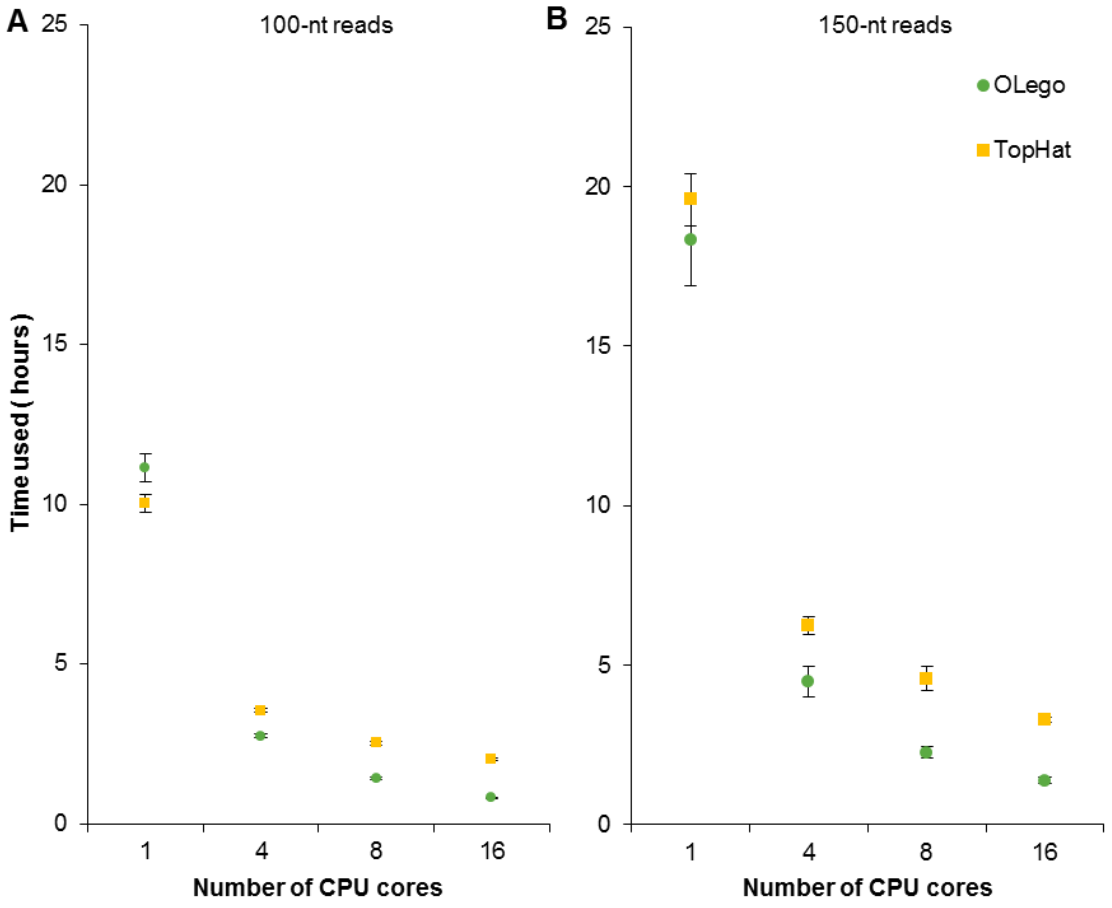


Figure 3. Comparisons of mapping speed. (A) Tests on 2×100-nt simulated reads; (B) Tests on 2×150-nt simulated reads. Running time (wall time) of TopHat (square) and OLego (triangle) on 10 million simulated paired-end reads with different numbers of CPU cores is shown. The values were averaged across three replicates for each test, with error bars indicating standard deviations.

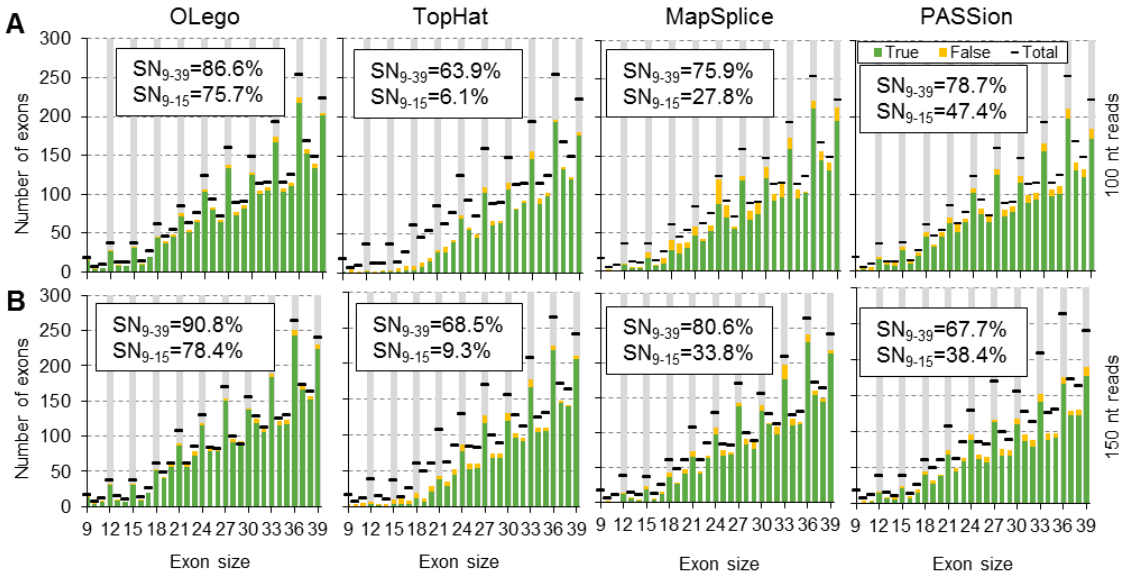


Figure 4. Discovery of small and micro-exons in simulated mRNA-Seq data. (A) 2×100-nt simulated reads; **(B)** 2×150-nt simulated reads. In each panel, internal exons within mapped reads were counted. The numbers of true (open columns) and false (solid columns) exons of different sizes, compared to the ground truth (horizontal bar) are shown for OLego, TopHat, MapSplice and PASSion, respectively. The overall sensitivity (SN_{9-39}) and the sensitivity for exons of size 9-15 nt (SN_{9-15}) are indicated on each plot.

Small or micro-exon discovery in simulated datasets

Alternatively spliced exons are generally shorter than exons that undergo constitutive splicing (69), and some are extremely small [e.g., 6 nt (70)]. Due to the limited information content encoded in such short sequences, micro-exons (<30 nt) (55) and their alternative splicing are intriguing with respect to their functional significance and underlying regulatory mechanisms. However, these exons are more likely to be missed in *de novo* searches, because they are much less likely to have seed sequences completely within the exon. Therefore, we evaluated the performance of OLego, TopHat, MapSplice and PASSion in finding small exons or micro-exons (9-39 nt). OLego consistently performed best in both sensitivity and specificity, compared to the other three programs (Figure 4). In terms of specificity, OLego achieved a PPV of 96.5% for 100-nt reads and 95.7% for 150-nt reads, respectively, compared to 91.7~93.6% for TopHat, 88.3%~93.3% for MapSplice, and 90.7%~91.8% for PASSion. OLego achieved an overall sensitivity of 86.6% for 100-nt reads and 90.8% for 150-nt reads, respectively, which was much higher than TopHat (63.9%~ 68.5%), MapSplice (75.9%~80.6%), and PASSion (67.7%~78.7%).

We further grouped exons according to their sizes to evaluate the sensitivity of each program. For exons of size 27-39 nt, MapSplice discovered a smaller number of exons than OLego (83.3% vs. 88.4% for 100 nt reads and 87.2% vs. 92.3% for 150 nt reads), and gave a lower PPV (91.3% vs. 96.6% for 100 nt reads, and 94.5% vs. 95.9% for 150 nt reads). TopHat and PASSion had the lowest sensitivity among the four programs (75.8% ~ 80.6% for TopHat, and 69.8%~80.1% for PASSion). Again, the advantage of OLego was most prominent in detecting extremely short micro-exons. For example, in detecting micro-exons of size 9-15 nt, OLego had a sensitivity of 75.7% for 100-nt reads and 78.4% for 150-nt reads, respectively, which was substantially higher compared to TopHat (6.1% and 9.3%, respectively), MapSplice (27.8% and 33.8%, respectively) and PASSion (47.4% and 38.4%, respectively).

Exon junction discovery in real data

After systematic evaluation of OLego using simulated data, we proceeded to analyze an mRNA-Seq dataset prepared from mouse retina RNA, which consisted of ~26 million 120 nt paired-end reads (68). We first examined known and novel exon junctions identified *de novo* by OLego. In total, mapping of these reads identified 234,440 unique exon junctions. Among them, 159,938 junctions (68.2%) were previously annotated, based on a comprehensive database of gene models derived from multiple sources (denoted as inclusive gene models; see Methods for more details) (53); more strictly, 137,606 (58.7%) junctions were annotated in RefSeq genes. We next binned all identified junctions according to the number of supporting reads, and categorized junctions in each bin into annotated junctions and three classes of novel junctions: novel junctions in which both splice sites are annotated separately, but the intron itself is not annotated (class I); novel junctions with only one site annotated (class II); and novel junctions with neither site annotated (class III) (Figure 5A). This analysis suggested that for exon junctions supported by >4 reads, 96.9% junctions were previously annotated, and an additional 0.69% exon junctions were class I novel junctions. On the other hand, for exon junctions supported by a single read, only 24.4% were previously annotated, and 39.8% were class III novel junctions. This trend is not surprising, because more abundant exon junctions are more likely to be known from previous data. As sequencing

depth increases, it becomes more likely to observe novel, rare splicing events, which are, however, complicated by sequencing and alignment errors.

Distinguishing novel exon junctions from artifacts introduced by alignment errors in real mRNA-Seq data is difficult. Nevertheless, we reasoned that there are two major sources of mapping errors that can introduce false exon junction detection. The first type is ambiguous determination of splice sites, due to repetitive sequences in double-anchor junction search; and the second type is errors introduced in single-anchor search when the number of matched nucleotides at the other end of the junction (anchor size) is limited. Manual examination of unannotated junctions suggested that the latter might be dominant. To study the relationship between anchor size and false junction detection, we binned all the aligned junction reads by the anchor size (Figure 5B). The rationale here is that the boundaries of mRNA/cDNA fragments in library preparation are random relative to the position of the splice sites, which is what we actually observed (Figure 5B, black curve). If all reads were aligned perfectly, the percentage of junction reads sampled from annotated (real) junctions should not vary as a function of anchor size. On the other hand, it is clear that a higher error rate is expected to occur when the anchor size is small, due to an increase in the chance of random matches. Therefore, examining the percentage of reads sampled from known junctions as a function of anchor size provides an independent method of estimating the bound of mapping errors. Specifically, with an anchor size >12 nt, 98.8% of splicing events in reads were mapped to annotated exon junctions, and an additional 0.23% were mapped to class I novel junctions. On the other hand, for junction reads with an anchor size of 8 nt, 86.4% of splices were mapped to annotated junctions, and an additional 0.3% were mapped to class I novel junctions. Therefore, we estimated that the false mapping rate of an exon junction read with anchor size of 8 nt could be as high as 12.3-13.3%, although these alignments represented a minor proportion of all junction alignments (1.86%). Similarly, the false mapping rate of an exon junction read with anchor size of 10 nt was estimated to be 1.9-2.9%. By requiring a more stringent anchor size of 10 nt, we identified 208,567 unique exon junctions, among which 76.4% were annotated previously in inclusive gene models. The proportion was 97.1% and 35.3% for junctions supported by >4 reads and by a single read, respectively.

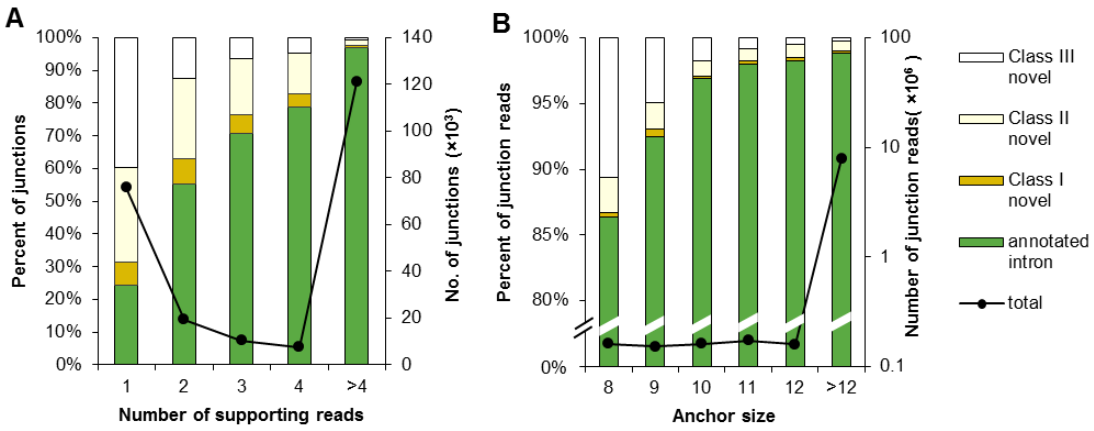


Figure 5. Distributions of exon junctions discovered in mouse retina mRNA-Seq data. . **(A)** The junctions found by OLego were binned according to the numbers of supporting reads. Different patterns indicate categories of junctions in the bar plot: annotated junctions; junctions with both splice sites annotated (Class I novel); junctions with only one splice site annotated (Class II novel); and junctions without any splice site annotation (Class III novel). The total number of junctions discovered in each bin is shown by the solid line with axis on the right. **(B)** The junction alignments were grouped according to their anchor sizes. The categories of the junctions are shown in the same way as in panel (A), and the numbers of junction alignments are shown by the solid line with the y-axis on the right.

Micro-exon discovery in real data

Our evaluation on simulated datasets suggested that OLego is particularly sensitive and accurate for micro-exon discovery. In this real mRNA-Seq dataset, we identified 1,665 micro-exons between 9 nt and 27 nt (Figure 6A and Supplementary Table 2), after requiring a minimal match of 10 nt at both ends for junctions flanking the micro-exon. Among these, 1,035 exons (62.2%) were annotated in inclusive gene models (53), and more restrictively, 715 (42.9%) were annotated in RefSeq genes. Among the remaining 630 exons that lack any evidence in current gene models, we examined the 5 ´splice site of the upstream intron and the 3 ´splice site of the downstream intron flanking each micro-exon. We found that 417 exons (66.2% out of 630 or 25% out of 1,665) had both the upstream and downstream constitutive splice sites annotated in the current gene models, as well as supporting reads that connect them to the micro-exon on both sides.

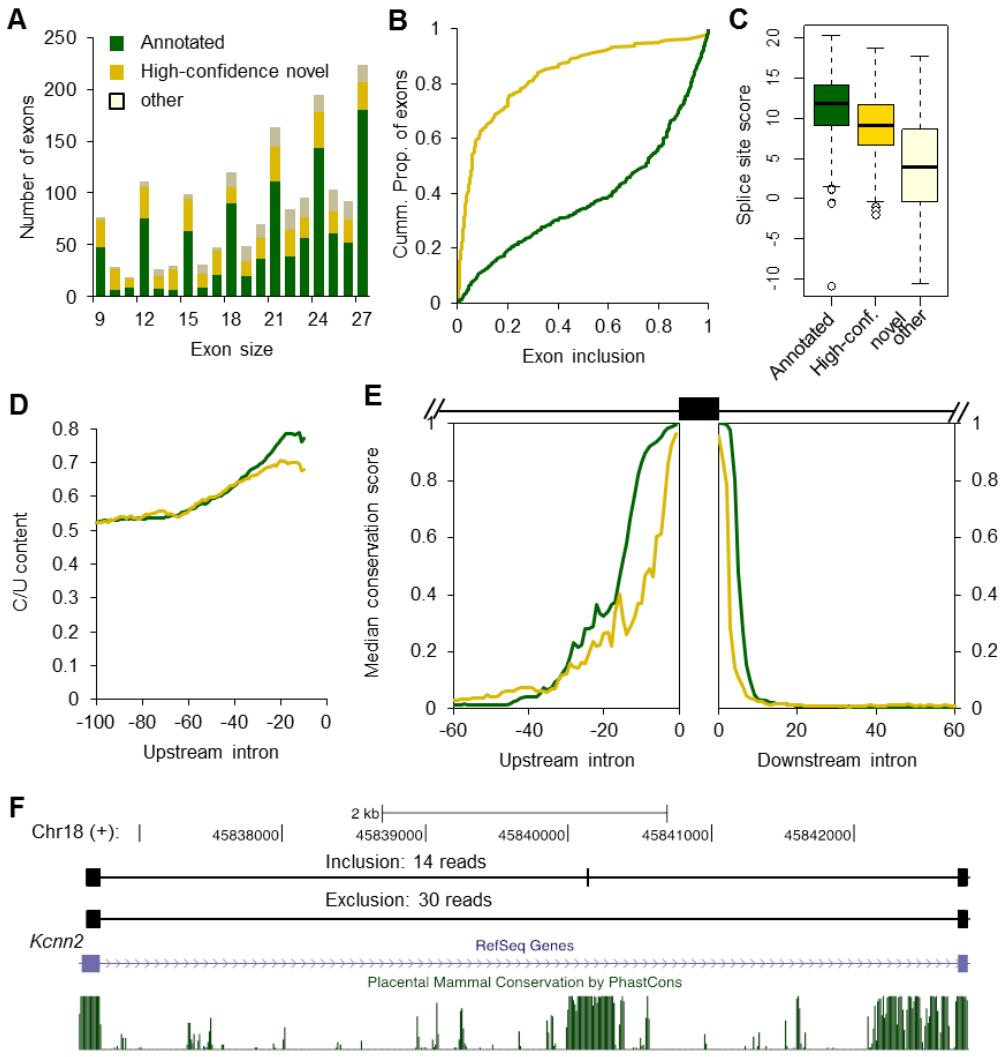


Figure 6. Discovery of micro-exons in mouse retina mRNA-Seq data. (A) Number of micro-exons identified by OLego. Exons are binned by their sizes (9~27nt), and in each bin, they are classified into three groups: annotated micro-exons in previous gene models (black), high-confidence novel micro-exons (exons with both flanking constitutive splice sites annotated; gray), and other (blank). **(B)** Cumulative distribution of exon inclusion level for annotated and high-confidence novel micro-exons; only those cassette exons with ≥10 reads that support either isoform were included for this analysis. **(C)** The distribution of total splice-site score (3 ´+ 5 ´splice sites) for each group of micro-exons is shown as a boxplot. **(D)** The pyrimidine (C/U) content in the upstream 100-nt intronic sequences, calculated using 10-nt sliding windows. **(E)** Cross-species conservation around the micro-exons. The medians of phastCons scores across 30 vertebrate species in the intronic regions immediately upstream and downstream of the annotated and high-confidence novel micro-exons are shown. **(F)** An example of a 9-nt novel micro-exon in the *Kcnn2* gene is shown. This exon is missing in current gene models (e.g., RefSeq) or cDNA/EST data (not shown), but both isoforms are abundant in the mouse retina (the two tracks on the top). The micro-exon is embedded in a longer stretch of conserved sequences.

This subset is expected to have a higher reliability, and we refer to it as high-confidence novel micro-exons. As a comparison, TopHat, MapSplice, and PASSion found 790, 713 and 1,242 micro-exons with the same criteria, respectively, among which 81, 85 and 163 exons are of high confidence with the same criteria (Supplementary Figure 1). Compared to OLego, very short micro-exons are under-represented in the results of all three programs, consistent with observations from simulation data.

A large proportion (988/1,665 or 59.3%) of the micro-exons have a size that is a multiple of three (Figure 6A). This is a prominent feature of regulated alternative splicing (71) and is consistent with our results on simulated data (Figure 4). Indeed, 42.2% (437/1,035) of annotated exons and 67.9% (283/417) of high-confidence novel exons are cassette exons, for which both inclusion and skipping were observed in these mRNA-Seq data. For these cassette micro-exons, novel exons tend to have much lower inclusion levels, compared to annotated exons (Figure 6B). The difference between annotated and novel exons can be explained in part by their difference in splice signals. Compared to the annotated micro-exons, novel exons have weaker summed 3´and 5´splice site scores (9.12 vs. 11.89, median; Figure 6C) and polypyrimidine tract (Figure 6D), although their scores are still clearly above background. Another not mutually exclusive possibility is that inclusion of novel cassette micro-exons shifts the reading frame more frequently, compared to annotated cassette micro-exons (47.7% vs. 16%), which would likely introduce premature stop codons (PTCs) and thereby trigger nonsense-mediated mRNA decay (NMD) to reduce the apparent inclusion level (72).

To assess the functional significance of the novel micro-exons, we examined their sequence conservation in vertebrate species (73). For both annotated and high-confidence novel micro-exons, we observed a high level of sequence conservation in flanking intronic regions (Figure 6E). For example, a 9-nt cassette exon in the *Kcnn2* gene is located in a long stretch of highly conserved sequences, and both isoforms were abundantly detected by mRNA-Seq, but not in previous cDNA/EST data (Figure 6F). Presumably, these regions harbor conserved *cis*-regulatory elements, which might be important for regulated splicing of these micro-exons and are thus under evolutionary selection pressure.

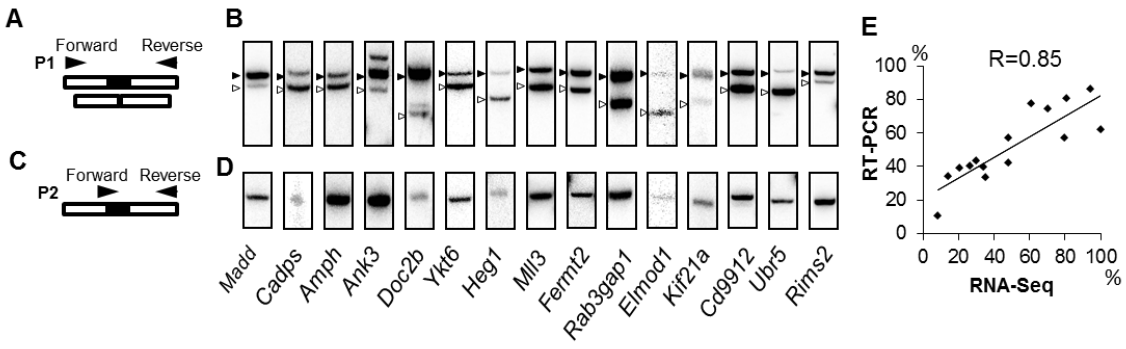


Figure 7. Experimental in vivo validation of micro-exons discovered by OLego. (A,C) Primers were designed either in the flanking exons to detect both micro-exon inclusion and skipping isoforms (A), or at the exon junction to specifically detect micro-exon expression (C). Primers positions and structure of each isoform are indicated (not to scale). **(B,D)** RT-PCR analysis of micro-exon expression in mouse retina using primers described in (A,C). Micro-exon included and skipped isoforms are indicated next to the corresponding bands by solid and empty arrowheads, respectively. **(E)** Correlation of micro-exon inclusion ratios estimated from mRNA-Seq data and those measured by radioactive PCR, as described in (B) (n=3).

In vivo validation of the micro-exons discovered by OLego

To assess the accuracy of OLego's micro-exon predictions, we experimentally tested the expression and inclusion ratios of micro-exons in mouse retina. We ranked the high-confidence novel micro-exons by the number of supporting reads for the inclusion isoform, and selected 15 exons for PCR validations (Supplementary Table 3 and Supplementary Figure 2). Two sets of primers were designed to validate each micro-exon: (i) primers were positioned in the flanking exons of the micro-exon to detect both exon inclusion and exclusion; and (ii) one of the primers was positioned on the exon junction spanning the micro-exon and a flanking exon, while the other primer was positioned in a flanking exon (Figure 7A, C; Supplementary Table 4). This ensured that we would both quantify the inclusion ratios and specifically detect the micro-exon, respectively. For all tested exons, we detected two isoforms with a size corresponding to the inclusion and exclusion of the micro-exon, respectively (Figure 7B). In addition, amplification with primers specific to the micro-exon junction confirmed the identity of the included/skipped micro-exon (Figure 7D). Therefore, OLego performs very well in micro-exon discovery, as we were able to validate15 out of 15 predicted novel micro-exons. This is further supported by the observation that the inclusion ratios estimated

from the RNA-Seq data and those measured in the PCR validation are highly correlated (Pearson correlation coefficient R=0.85; Figure 7B, E, Supplementary Table 3).

Discussion

Here we present OLego, a program designed for fast mapping of hundreds of millions of mRNA-Seq reads to the reference genome with high specificity and sensitivity, which allows identification of known and novel exon junctions. Since the first publication of mRNA-Seq studies (19), the technologies have evolved very rapidly, with the most prominent features including increases in read length and throughput, and a reduction in sequencing errors.

The first generation of tools that align RNA-seq reads to the genome is based on the construction of a database of known or predicted exon junction sequences (8,74,75), so that junction reads can be mapped against this junction sequence database without alignment gaps. This strategy is fast and accurate for alignments to annotated exon junctions. However, it relies on the fact that reads are short (~36 nt), so that they rarely span more than one junction, and another important caveat is that this approach does not allow the discovery of exon junctions *de novo*. As the read length increases, it becomes more and more frequent for a read to span three or more exons, but it is difficult to build a sequence database of alternative isoforms that span many exons, while preserving the "uniqueness" of sequences that can potentially match mRNA-Seq reads. Algorithms designed specifically to map spliced mRNA-Seq reads were subsequently developed, with TopHat being one of the first (76), followed by several others, such as SpliceMap (35), GSNAP (38), MapSplice (33) and PASSion (34).

Although different heuristics were employed in each of these programs, most of them use a seed-and-extend strategy, which was also used in programs developed earlier to map traditional cDNA/EST sequences to genomic DNA sequences, such as sim4 (77), BLAT (57) and exonerate (78). With this strategy, the size and position of the seeds are critical determinants of mapping sensitivity. In general, a match of at least one seed in each exon is critical for successful alignment of a read, although tricks like single-anchor junction search can be used to match sequences near the ends of a read. To achieve sensitivity, these earlier programs typically used short seeds of size 11-12 nt. BLAT is

one of the first programs to allow fast mapping of cDNA/EST sequences to the whole genome, by hashing the whole genome using non-overlapping seeds or tiles (default=11 nt). However, speed is a bottleneck in processing ultrahigh-throughput mRNA-Seq data.

To improve the speed of genome-wide mapping of large numbers of short reads, various schemes have been used to index the reference genome sequences to enable faster querying. For example, GSNAP uses a hash table indexing all the k-mers (k typically in the range from 11 to 15 nt) every 3 nt in the reference genome. The overlapping 3-nt spaced seed hashing scheme is necessary to reduce the memory footprint to around 4 GB. An alternative approach to hashing is the BWT- and FM-index-based method employed by many programs, including Bowtie, which is integrated into TopHat and MapSplice. This invertible full-text indexing scheme is more memory-efficient, and allows fast query of sequences of varying length, in contrast to the fixed size of seed sequences in hashing-based methods. This flexibility makes it possible to align different types of reads with different granularity, e.g., fast alignment of exon-body reads without requiring seed partitioning, followed by alignment of spliced reads using short seeds, and very short or micro-exons of varying sizes.

Most currently available mRNA-Seq read splice-mapping tools typically segment reads into non-overlapping, relative long (~25 nt) seeds, which are mapped to the genome without gaps by an external mapper. The relatively long seeds restrict the number of hits, so that the temporary results generated by the external mapper are manageable in post-processing steps to produce final junction alignments. However, even with relative long seeds, the pipeline-based methods still generate temporary files of enormous size, which can be a significant concern regarding both space and speed when these files are parsed to produce final results. For example, with the basic configuration (paired.cfg), MapSplice required about 140 and 200 GB of disk space to store temporary files to align the 10 million paired-end 100- or 150-nt reads, respectively, in our simulation. In the more exhaustive mode (Try_hard.cfg in the package), the disk usage of MapSplice increased to over 500 GB for 100-nt reads and 800 GB for 150-nt reads, respectively. Interestingly, with this mode we did not observe an increase in sensitivity, but did observe a dramatic drop in accuracy (data not shown).

The relatively long seeds increase the chance that the seeds themselves span exon junctions, reducing the number of seeds mapped to exonic sequences, which are critical for final sensitivity. For example, the median size of mammalian exons is ~120 nt. In this case, ~21% (25/120) of 25-nt sliding windows overlap with exon junctions. This problem gets worse for alternative exons, especially those regulated to have variable inclusion levels in different conditions, such as different tissues. For example, the median size of cassette exons regulated by the neuron-specific splicing factor Nova is ~80 nt (79), so that ~31% (25/80) of 25-nt sliding windows overlap with exon junctions. Smaller seeds greatly reduce the chance of overlaps with exon junctions in seed sequences. With the ~14-nt seeds used in OLego, a read covering an exon of ≥28 nt is guaranteed to have at least one seed inside the exon, which increases the sensitivity of mapping reads sampled from these relatively small exons. Instead of using a smaller seed size, PASSion (34) uses a different strategy, based on a pattern-growth algorithm. This method can only be applied in paired-end mRNA-Seq data when one exonic read is aligned in the first pass without allowing large gaps, while the other read, which spans one or two exon junctions, is missed in the first pass and is to be refined in later steps. In this scenario, the aligned read in a pair is used as an anchor, and local searches of the (maximum) unique substrings starting from the ends of the other read are performed using a pattern-growth algorithm, constrained by the maximum intron size. This process can continue iteratively until the substrings cover the whole read. The advantage of PASSion in eliminating the read-segmentation step is attractive, and this algorithm was reported to show competitive performance compared to several other programs. However, it is unclear how this algorithm handles sequencing errors, the repetitive nature of substrings, and longer mRNA-Seq reads in which both reads in a pair span exon junctions. In practice, OLego achieved both higher sensitivity and accuracy in discovery of exon junctions and micro-exons using realistic simulated data.

OLego aligns each read independently in one pass, without filtering of junctions based on their summary statistics derived from all reads, such as the uniformity of the positions of reads mapped to the junction used by MapSplice and the read coverage around the junction used by TopHat and PASSion. To ensure the accuracy of junction mapping, OLego limits the *de novo* search to canonical GT/AG splice sites, and uses a

built-in model of exon-junction strength that combines splice-site scores and intron size. This intentional choice is due to the fact that canonical splice sites account for ~99% mammalian introns (66). The exon-junction scoring is effective to find the real splice site in a single-anchor junction search, in which multiple hits flanked by splice-site dinucleotides can be found when the anchor sequence is short. It also improves the accuracy in a double-anchor junction search, when ambiguity exists due to repetitive sequences near the splice sites. As a result, OLego achieved an accuracy comparable to that of TopHat and better than those of MapSplice and PASSion, while at the same time having a much lower FNR. The advantage of OLego compared to the other programs is particularly prominent for exon junctions of low abundance, as demonstrated in simulations. Nevertheless, the default parameters are chosen to balance accuracy and sensitivity, which is suitable for many applications of mRNA-Seq to quantify splicing levels. For efforts focusing on the discovery of novel junctions, including those of low abundance, filtering of junctions based on supporting evidence and anchor size can certainly improve the accuracy further.

While this paper was under revision, another program named TrueSight was published (80). TrueSight also used splice site motifs together with other features to build a regression model to distinguish true vs. false positive exon junctions, and reported improved PPV and sensitivity compared to TopHat, MapSplice and PASSion. One difference between TrueSight and OLego is that the former builds exon junction models on the fly, using junction reads already mapped, and updates the model and the alignment iteratively using an EM algorithm. The benefit of the EM algorithm is not very clear, given that a large number of exon junctions from reads mapped in previous steps (or in annotated gene models) is already available, and logistic regression is in general not very sensitive to some noises. In our experiment, we were able to use 10% of training data to derive our logistic regression model and obtain essentially the same results (data not shown). On the other hand, the iterative procedure in TrueSight appears to be computationally expensive, so that TrueSight is significantly slower than all the other programs the original authors compared and has a relatively large memory footprint (10Gb memory per 30 million reads). Another important limitation of TrueSight is that

it also relies on an external mapper for seed mapping, sharing the same limitation on seed size as TopHat and MapSplice, for which OLego aims to improve.

We employed several strategies to achieve fast mapping with small seeds to the mammalian-sized genome. First, we require perfect matches in seed sequences, given the fact that sequencing errors in typical mRNA-Seq data are as low as 0.5%. Mismatches, including substitutions and indels, are handled when the alignments are refined for each exon. Second, after hits of seeds are clustered, candidate alignments are ranked and filtered by the number and uniqueness of matched seeds. Therefore, the later time-consuming steps to locate exon junctions are only applied to the most promising candidate alignments. Third, BWT- and FM-index-based querying in the genome is not only applied in the step of seed mapping, but also in the later steps to locate splice sites (single-anchor junction search) and micro-exons. The capability of fast querying of sequences of different sizes using BWT is particularly helpful. Finally, we do not need to filter the alignments according to the abundance of each junction, so that each read can be mapped independently in one pass. This makes it possible to support multiple threading in the whole cycle of alignment. Indeed, although OLego performed more exhaustive searches than TopHat, the speeds of these two programs were still comparable, even with a single CPU core. Furthermore, the speed of OLego increased faster than that of TopHat as the number of CPU cores increased, such that with ≥8 CPU cores, OLego used half or less time, compared to TopHat. The other two programs, MapSplice and PASSion, were substantially slower in our comparison. We estimate that on 8 CPU cores, OLego can map a typical lane of 200 million paired-end mRNA-Seq reads of 100 nt and 150 nt to the mammalian genome in ~29 and ~46 hours, respectively. Combined with its small memory footprint, OLego can efficiently run on desktop workstations. It is also worth noting that increasing the seed size will further improve OLego's mapping speed, despite the risk of potential decrease in sensitivity of exon junction detection, especially for those flanking small or micro-exons.

We paid special consideration to searches of very small or micro-exons. Even with the small seeds used in this study, these exons might still lack internal seed sequences without overlap with exon junctions. However, these exons can be recovered when matches to sequences in the flanking upstream or downstream exons are found, and

the micro-exon sequences can be determined accordingly, so that they can be effectively searched against the indexed genome together with the flanking splice sites. As demonstrated by simulation, OLego was successful in identifying most of the extremely small exons of size 9-15 nt (75.7- 78.4%), whereas TopHat and MapSplice missed most of them (only 6.1-33.8% identified). PASSion identified more exons in this range (38.4%~47.4%) than TopHat and MapSplice, but the numbers were still much smaller than OLego's. TopHat provided an optional "micro-exon-search", which is supposed to improve the sensitivity of micro-exon search. However, even with this option enabled, TopHat only found 12% of these extremely small exons in the 100-nt dataset, as compared to 75.7% by OLego. Finally, we were able to identify over 400 high-confidence novel micro-exons in a single mRNA-Seq library of moderate depth (~26 million paired-end reads) prepared from mouse retina RNA. The inclusion level of these exons is lower than that of annotated ones, which is likely why they were not previously identified, and this can be explained by their weak splicing signals. However, we were able to validate 100% of the novel micro-exons tested by RT-PCR, demonstrating OLego's high sensitivity and accuracy. Some of these micro-exons likely have functional significance, as judged from their deep phylogenetic conservation (see also Supplementary Figure 2).

With its high sensitivity and accuracy and fast mapping speed, OLego can be used for efficient alignment of large-scale mRNA-Seq data being generated at unprecedented rate and depth. It can be combined with downstream analysis tools for transcript reconstruction and quantification to facilitate the process of revealing the transcriptomic complexity of mammals and other species.

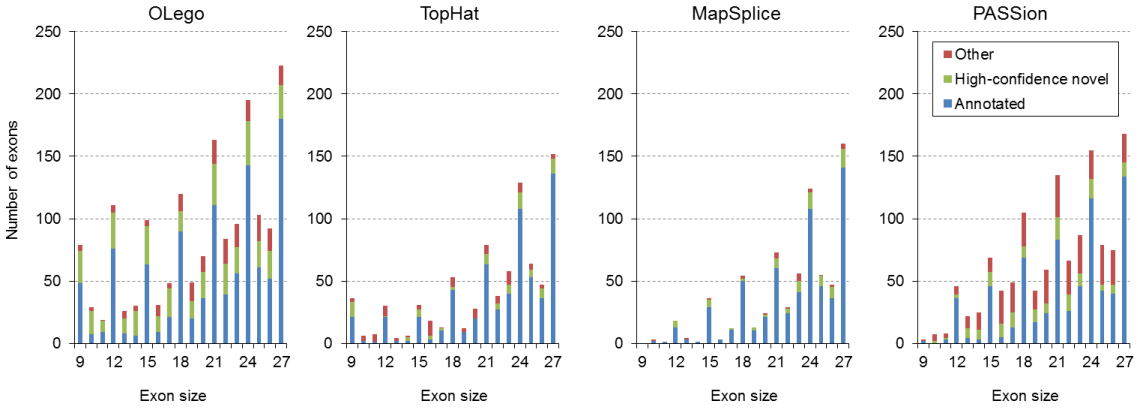
Acknowledgements

We would like to thank Michael Schatz and Martin Akerman for critical reading of the manuscript, and members of the Krainer, Zhang, and Robert Darnell labs for helpful discussion. C.Z. would also like to thank the Darnell lab for computing resources and support.

Funding

This work was supported in part by National Institutes of Health (GM74688 to M.Q.Z. and A.R.K., K99GM95713 to C.Z.); and National Basic Research Program of China (2012CB316503 to M.Q.Z.).

Supplementary Figures



Supplementary Figure 1. Number of micro-exons identified by OLego and the other three programs. Exons are binned by their sizes (9~27nt), and in each bin, they are classified into three groups: annotated micro-exons in previous gene models (blue), high-confidence novel micro-exons (exons with both flanking constitutive splice sites annotated; green), and other exons (red).

Supplementary Figure 2. UCSC genome browser screenshots of the validated micro-exons. Structures of the inclusion and exclusion isoforms are shown in the user supplied track. The number of supporting reads is indicated to the left of each isoform. Known transcript structures (UCSC genes, RefSeq genes and ESTs) are also shown in the screenshots. The micro-exons are in the same order as they are listed in Supplementary Table 2.

(There are 15 figures so they are not included, please find at NAR online

http://nar.oxfordjournals.org/content/41/10/5149/suppl/DC1)

Supplementary Tables

Supplementary Table 1. Details of the coefficients obtained from logistic regression.

Species		Estimate	Std. Error	z value	Pr(>|z|)
mm9	Intercept	1.13E+00	1.26E-02	89.11	<2e-16
	Intron size	-4.81E-05	2.79E-07	-172.15	<2e-16
	Splice-site score	2.77E-01	1.37E-03	202.28	<2e-16
hg18	Intercept	1.63E+00	1.16E-02	141.4	<2e-16
	Intron size	-5.35E-05	2.89E-07	-185	<2e-16
	Splice-site score	2.18E-01	1.14E-03	191.9	<2e-16

Regression were done using glm() in R, z values were given by Wald statistic.

Supplementary Table 2. Micro-exons identified in mouse retina RNA-seq data.

(This table is not included due to its size, please find at NAR online http://nar.oxfordjournals.org/content/41/10/5149/suppl/DC1)

Supplementary Table 3. List of the validated novel micro-exons.

Gene symbol	Exon size (nt)	Upstream exon pos	Exon pos	Downstream exon pos	Strand	In reads	Ex reads	In%
Madd	9	chr2:91013270-91013350	chr2:91012484-91012492	chr2:91010842-91010904	-	270	63	
Cadps	15	chr14:13282236-13282383	chr14:13274366-13274380	chr14:13273355-13273429	-	60	115	
Amph	9	chr13:19192441-19192523	chr13:19193771-19193779	chr13:19194741-19194879	+	46	108	
Ank3	12	chr10:69390443-69390566	chr10:69392195-69392206	chr10:69395157-69395236	+	42	27	
Doc2b	27	chr11:75599595-75599674	chr11:75599019-75599045	chr11:75595123-75595197	-	32	2	
Ykt6	15	chr11:5859309-5859382	chr11:5860430-5860444	chr11:5861191-5861291	+	29	113	
Heg1	24	chr16:33720735-33721088	chr16:33722365-33722388	chr16:33725511-33725729	+	29	53	
Mll3	23	chr5:24810496-24810559	chr5:24809915-24809937	chr5:24808524-24808756	-	29	31	
Fermt2	21	chr14:46084399-46084620	chr14:46082390-46082410	chr14:46081791-46081915	-	28	7	
Rab3gap1	21	chr1:129835623-129835725	chr1:129838171-129838191	chr1:129838951-129839155	+	27	29	
Elmod1	24	chr9:53772207-53772275	chr9:53771749-53771772	chr9:53769130-53769180	-	25	149	
Kif21a	12	chr15:90779437-90779475	chr15:90778755-90778766	chr15:90774196-90774332	-	23	0	
Cd9912	18	chrX:68682222-68682341	chrX:68678590-68678607	chrX:68677199-68677264	-	23	63	
Ubr5	27	chr15:37900017-37900102	chr15:37898682-37898708	chr15:37898093-37898260	-	21	232	
Rims2	12	chr15:39123648-39123858	chr15:39137638-39137649	chr15:39176856-39177166	+	19	8	

Supplementary Table 4. Primers designed to detect the micro-exons and the splice events.

Gene Name	Primer Sequence 5'-->3'				PCR product length for specific primer pairs (nt)		
	Micro-exon flanking primer		Micro-exon junction specific primer		F1/R1		F2/R1 or F1/R2
	F1	R1	F2	R2	exon inclusion	exon skipping	exon inclusion
Madd	CCACCAATGCAGAAGTGCTA	TTAAGAATGGGCTGGGTGTT		GGCTAAGGCCTTCAGGAACT	121	112	89
Cadps	CCCACAGTCAATATGCACCA	CTTTGCCACCAAAAGTGTGA	CCAAGAGTTTGCTAAAGAGTGGC		178	163	97
Amph	GTGATGACGAAACTCGGTGA	TGCAGGTGCTAATGTGTGGT	CCAGAGCGAGGTGTGACTC		174	165	116
Ank3	TCACAGGGGACACTGACAAG	CCTTGCGTAGGAGCTTCTGT		CCGAACTTATCTTGGGACTGG	167	155	129
Doc2b	ACGCTGGACTTCAGTCTGCT	CAGGCAGCAGGTGTAGTTTG	CAAGGTGCCAAAGCTGATG		160	133	95
Ykt6	CAAGTCAACTGATTGTGGAACG	AGGCCACTCTGGAAGGGTAT	GAACAAGCTCCATAAACAAGAGAG		159	144	106

Heg1	GTGGGAGGAGTTACGCAGAG	GACATTGCAGACGTGTGAGG	GAGCATTCTCTGCTTCACCG		205	181	122
Mll3	AAATTTTGGTCCAGGCTTTG	GCTGTGCGCTGTTTAGCTG	CTTTGTCAAAACCAGAAAGGC		185	162	170
Fermt2	CCGCGGTACCTGAAGAAGTA	AGTGTGTGATGCCGAACTCA	GAGCAAACAGCCAGGCTG		169	148	146
Rab3gap1	CTGGAGCAGCCTGAGGTATC	CTGGGCAGCCTGATTTCTTA	CTCAGCGGCTGACTGAATC		160	139	81
Elmod1	CTCTCCGACTCTGTCCATCC	CCAATGGCTTTATCCATCTTTT	CATCCAAAATGCAGGGATATC		104	80	89
Kif21a	ACTCCTCTTTGTCCGAGGTG	GGTGTGCCCTTCAGCTATGT	GTGCACAGATCCACCAGAAG		119	107	102
Cd9912	GACAGAAACTGGCACCATTG	CCCCTTCACGTAGTCTGCAT		CTGCATTGAGGCCCTCTG	160	142	146
Ubr5	GCTGAGAAGCTCCTCCAGTTC	TGTTGGTCATCTGGTGGTCT	GACCTGCCTCCCTTGCTC		200	173	128
Rims2	GCAAGAGCAGAAGGGTGATG	TTCTTGTTGTTTTCGGCACA		CCACTTTGTCCTCCTTGTTTG	193	181	155

Chapter 3 :

Identification of Exons and Splicing Events from Alignment Results

Introduction

An important feature of RNA-Seq is its ability to sequence the transcriptome at an unprecedented depth, hence makes it possible to observe rare transcripts which cannot be observed previously with traditional techniques (like EST). This also enables us to discover novel structure elements from the data, including novel exons and junctions, and furthermore, novel AS events.

Transcriptome reconstruction is the process to find the structures of transcripts expressed in the data. There are many tools developed in the last few years to reconstruct transcriptome from RNA-Seq experiments, as described in Chapter 1. Since AS events are variance of local transcript structures, a straightforward approach to recover AS events from RNA-Seq data is to extract the information from the results of methods like Cufflinks or Scripture (41,42).

However, unnecessary complexity has been introduced by this indirect approach, because AS events are local structural alternations which can be identified with local evidence directly without knowing the full structures of the transcripts. The transcriptome reconstruction itself is a complex problem. On one hand, the reads from

RNA-Seq data are relatively short, and only cover short regions in the transcripts, giving little evidence of the full structure. On the other hand, the complexity of the transcriptome and the noise from pre-mRNA make the reconstruction even harder, leading to a number of non-identifiable gene models (81).

To reduce the complexity, some of the transcriptome reconstruction programs only report the minimum set of isoforms which can explain the reads at a gene locus. Although this strategy can presumably remove some noise, it discards some signals and results in a low sensitivity on rare AS events.

In our study, instead of using transcriptome reconstruction programs to identify AS events, we use in-house scripts to connect the evidences in the data, e.g., junction reads, to recover AS events. We did a simulated test to benchmark the performances of different programs when identifying exons from RNA-Seq data. And we demonstrated how many exons could be missed by traditional transcriptome reconstruction methods, and "simple connection" method is able to recover more exons than these transcript-based methods.

Methods

Simulated dataset

We generated simulated mRNA-Seq reads using the program BEERS from the RUM package (53). For mouse (mm9), BEERS uses gene models derived from 11 annotation tracks (AceView, Ensembl, Geneid, Genscan, NSCAN, Other RefSeq, RefSeq, SGP, Transcriptome, UCSC, and Vega) in the UCSC genome browser to avoid bias towards or against any particular set of gene annotations. 30,000 gene models were selected randomly from the pool and 10 million 100-nt paired-end reads were sampled, with an expression profile which mimics patterns in real data.

Recovery of the exons with the programs

The simulated reads were first mapped to reference genome sequence (mm9) with OLego (39) using default options. Then Cufflinks (1.3.0) and Scripture (beta2) were run to

reconstruct the transcripts without any transcript annotation. For Cufflinks, -F=0 and --min-frags-per-transfrag=0 were used to output all isoforms it could find. While for Scripture, paired end mode was used and only segmentation task was done to re-construct the transcriptome. Afterwards, the exons were extracted from their results to evaluate their performances.

Simple connection method

To recover the exons directly from the read mapping results without transcript reconstruction, we first identified the reliable set of junctions by requiring at least 3 supporting reads on each junction. Then we connect nearby 3' and 5' splice sites to identify exons, allowing a maximum exon length up to 1000 nt. If there is potential intron in the region of an exon (supported by junction reads), we check the coverage of the "intron" and only report the exon if 90% of this "intron" is covered, which might indicate an intron retention event. BEDTools was used in this step(82). Since the transcript start and end sites cannot be interpreted by junction reads, known transcript boundaries were used as reference.

Results

Comparison of the exon recovery

We binned the exons identified by different methods according to their coverage in the simulation and compared the numbers between the methods, shown in Figure 1. All the approaches had low sensitivity for lowly covered exons. For exons with higher coverage, Cufflinks and Scripture had a maximum discovery rate around 60%, and Scripture had a better sensitivity than Cufflinks, especially for exons with medium coverage. The simple connection method did best at all coverages, and was able to identify nearly 100% of the highly covered exons in the simulated dataset. These results showed that transcript-based methods have a bottle neck when they are used to identify exons, while a simpler method based on local structures can outperform in this specific task. It is also striking to see that the simple connection method achieved a comparable or higher accuracy compared to the other two methods.

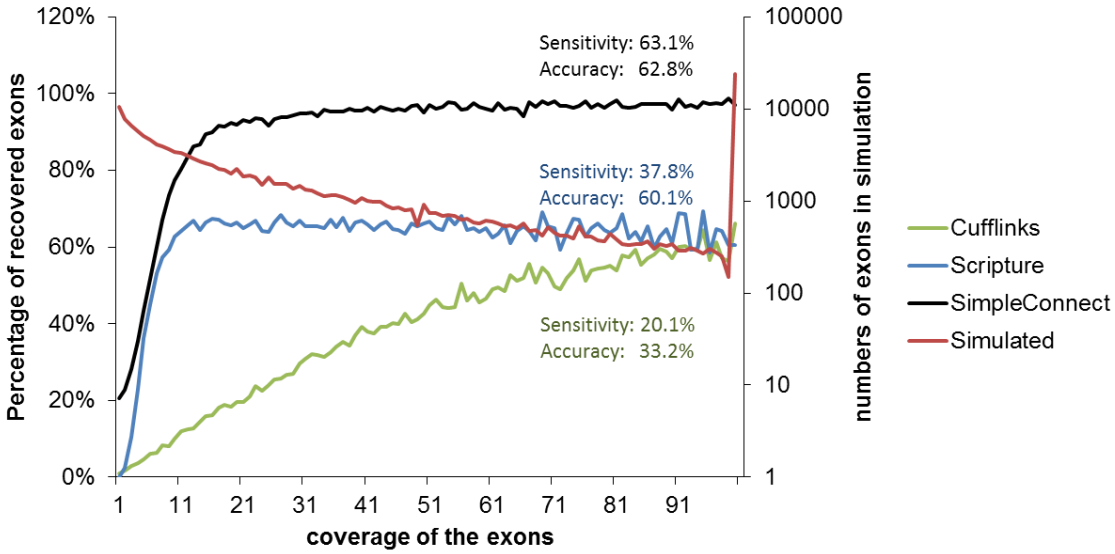


Figure 1. Exon recovery by different approaches. Exons were binned according to their coverage in the simulation (x-axis). Percentage of exons recovered from the data (left y-axis) were shown in different colors for Cufflinks, Scripture and SimpleConnect (the simple connection method). Total numbers of exons in each bin (right y-axis) are shown by the black curve.

Discussion

With a simulated test, we showed that a simple connection method has advantage on identifying exons from RNA-Seq alignment results. Compared to transcript-based methods like Cufflinks or Scripture, it demonstrated better sensitivity on exon discovery and comparable prediction accuracy.

The method proposed was initially developed to extract splicing events from alignment results from RNA-Seq data. This is to generate a database for later quantification of the events with SpliceTrap(10) (Chapter 4). Although there are potentially a large number of false positives in the recovered exons, we can further filter out those unreliable exons with SpliceTrap, which has a built-in dynamic cutoff model based on both exon coverage and exon length.

The proposed method here is local structure based, which grantees a much faster speed compared to transcript based methods. This is due to its low complexity and not requiring any statistical model.

Chapter 4 :

SpliceTrap: a Method to Quantify Alternative Splicing under Single Cellular Conditions

Abstract

Alternative Splicing (AS) is a pre-mRNA maturation process leading to the expression of multiple mRNA variants from the same primary transcript. More than 90% of human genes are expressed via alternative splicing. Therefore, quantifying the inclusion level of every exon is crucial for generating accurate transcriptomic maps and studying the regulation of alternative splicing.

Here we introduce SpliceTrap, a method to quantify exon inclusion levels using paired-end RNA-seq data. Unlike other tools, which focus on full-length transcript isoforms, SpliceTrap approaches the expression-level estimation of each exon as an independent Bayesian inference problem. In addition, SpliceTrap can identify alternative splicing events under a single cellular condition, without requiring a background set of reads to estimate relative splicing changes. We tested SpliceTrap both by simulation and real data analysis, and compared it to state-of-the-art tools for transcript quantification. SpliceTrap demonstrated improved accuracy, robustness, and reliability in quantifying exon-inclusion ratios.

SpliceTrap is a useful tool to study alternative splicing regulation, and can be used in combination with full-transcript quantification tools, to generate high-resolution transcriptomic maps.

Availability and implementation: SpliceTrap can be implemented online through the CSH Galaxy server http://cancan.cshl.edu/splicetrap and is also available for download and installation at http://rulai.cshl.edu/splicetrap/.

Introduction

In higher eukaryotes, a given transcribed locus can generate several mature mRNA isoforms via the process of alternative splicing (AS). AS is frequently a regulated mechanism that coordinates the removal of the internal non-coding portions of the transcripts (introns) with the differential joining of the coding and 5'/3' untranslated portions (exons). As a result, proteins with similar, different, or antagonistic activities can be generated from a single genomic locus (83,84). In addition, AS can lead to downregulation of gene expression by diverting some of the mRNA isoforms to the nonsense-mediated mRNA decay pathway (17).

More than 90% of human genes express primary transcripts that undergo AS (8,9). Owing to the regulatory power of this process, an increasing number of studies are being directed at understanding AS regulation at the single-exon level (17,85-87). In general, researchers in the splicing regulation field have utilized comparative approaches to reveal tissue-specific (88,89) or disease-related (90) AS events. However, such methodologies have not been used to generate maps of AS activity within one cellular condition. The completion of such maps would add a higher level of resolution to transcriptome analysis, allowing precise quantification of exon inclusion levels within a population of related isoforms.

Until recently, systematic analysis of AS was done using expressed sequence tags (EST) (91-93) or specialized microarrays (8,85,86,94). These techniques facilitated the discovery of a large number of alternative transcripts, and the extraction of distinctive features of alternatively spliced exons. Nevertheless, these techniques suffer from several

limitations. ESTs are subject to cloning biases—especially towards the 3' end of transcripts—low coverage, and insufficient robustness to allow reliable quantification. Likewise, the specificity of splicing microarrays is negatively affected by cross-hybridization with related mRNA molecules.

The development of deep-sequencing technologies provided an alternative to ESTs and microarrays for transcriptomic quantification. Recent studies by Pan et. al. and Wang et. al. utilized single-end RNA-seq to analyze a series of human tissues. In Pan et. al., the inclusion level of alternative exons was quantified as the % of the number of reads that match the two splice junctions formed by exon inclusion, over the splice junction formed by exon skipping. Wang et. al. also utilized splice-junction reads for quantification of minor isoforms with different frequencies, as a function of the read coverage or RPKM (reads per kilobase of exon per million mapped reads) (9). Although both studies demonstrated improved coverage relative to microarrays and ESTs, they utilized only isoform-specific reads, leaving out the majority of reads, which map to common exons of different isoforms.

An improved version of the deep-sequencing technique utilizes paired-end tags (95), which allows a significant gain of coverage and a reduction in read ambiguity through the generation of linked tag pairs that span longer stretches of sequenced template. This technology is especially suitable for AS profiling, because many exon-mapped tags are expected to span splice junctions, and these can be exploited to improve AS quantification.

Two recent methods exploit paired-end sequencing information for transcript quantification: Cufflinks (32), which is based on a previous RNA-seq model for single-end reads (96) and Scripture (42). Both can reconstruct transcript structures using directed graphs, and assign FPKM (fragment per kilobase of exon per million mapped reads) or RPKM values to every transcript, without relying on a reference genome. Cufflinks uses a rigorous mathematical model to identify alternatively regulated transcripts at each locus. Scripture employs a statistical segmentation model to distinguish expressed loci, and filters out experimental noise. Both methods perform very

well in identifying and quantifying full transcript levels, and can be used to deduce isoform changes between two cellular conditions. However, neither method aims to quantify the inclusion ratio of every exon under single cellular conditions, as required to generate maps of AS regulation.

Here we introduce SpliceTrap, a method to quantify exon inclusion levels in paired-end RNA-seq data. SpliceTrap generates alternative splicing profiles for different splicing patterns, such as exon skipping, alternative 5' or 3' splice sites, and intron retention. SpliceTrap utilizes a comprehensive human exon database called TXdb (see Methods) to estimate the expression level of every exon as an independent Bayesian inference problem.

We tested SpliceTrap both by simulation and real data analysis. Compared to Cufflinks and Scripture, it demonstrated improved accuracy, robustness, and reliability in quantifying splicing activity. In summary, SpliceTrap is suitable for studying AS patterns and constructing high-resolution transcriptomic maps, when used in combination with gene-expression profiling tools. SpliceTrap can be implemented online through the CSH Galaxy(97) server http://cancan.cshl.edu/splicetrap and is also available for download and installation at http://rulai.cshl.edu/splicetrap/.

Methods

Database construction

To quantify exon-inclusion levels, we designed an exon-trio database called TXdb. First, we captured all known transcripts encoded by every human gene (Figure 1A,B), using annotations from RefSeq (98) (downloaded from the UCSC genome browser, hg18) and the EST-based AS database dbCASE (99). Second, to account for every possible exon-skipping event, we subdivided each transcript set (i.e., encoded by the same gene) into exon trios, by sliding a 3-exon window along the transcript (Figure 1C). In particular cases in which an exon was flanked by more than one assembly of flanking exons, every possible combination was represented in TXdb as a separate case. About 20% of the exons in TXdb are represented by more than one assembly of flanking exons (i.e., trios or

duos). The pie charts in Supplementary Figure 1 show the types and numbers of exons represented by one or multiple assemblies.

Next, we formatted the database to allow quantification of exon skipping (CA: cassette exon). We assumed that the middle exon was a cassette exon (E2) (regardless of whether it is annotated as alternatively or constitutively spliced) and the flanking exons

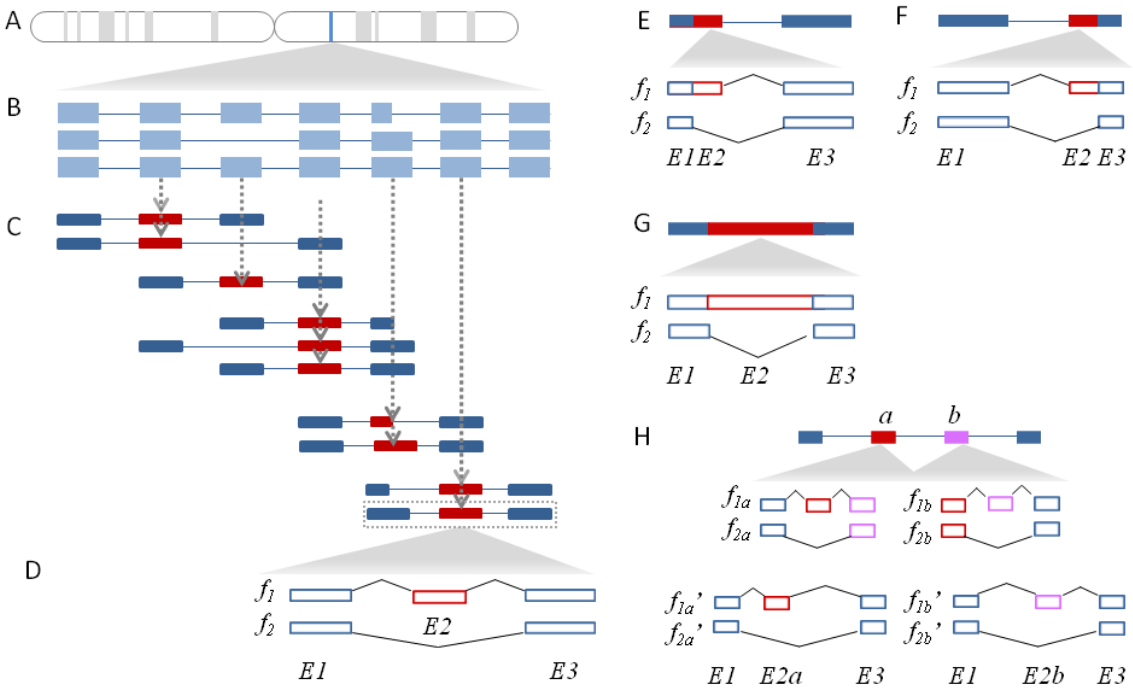


Figure 1. TXdb assembly. (A) From a given gene expression locus (blue strip) (B) we extracted all the known transcript isoforms using available transcriptome annotations. (C) Using a 3-exon sliding window, we subdivided the transcript isoform population into exon trios, accounting for all known transcriptomic variability. Every exon trio is used as an independent mappable unit, wherein the middle exon (red block) is queried for alternative splicing activity and the flanking exons (blue blocks) are treated as constitutive exons. (D) Two isoforms are constructed for each trio. Every exon-skipping event is represented by an inclusion isoform (f_1) and a skipping isoform (f_2) which comprise a pair of flanking exons (E1, E3) and an alternative exon (E2) present in f_1 but not in f_2. To examine additional types of alternative splicing, such as (E) alternative 5' splice sites, (F) alternative 3' splice sites, and (G) intron retention, we generated exon duos to compare extended isoforms to shortened isoforms. (H) SpliceTrap can detect consecutive alternative exons. When the alternative exon *a* or *b* is used as a flanking exon in an exon trio (f_{1a}, f_{2a}, f_{1b} and f_{2b}), if it is skipped, the exon trio will not pass the coverage cutoff, and thus will not be considered to be reliable. However, if substitute exon trios are present in TXdb (f_{1a}', f_{2a}', f_{1b}' and f_{2b}'), when $f_{1a}' > f_{2a}'$ and $f_{1b}' < f_{2b}'$ or vice-versa (referring to their expression levels), exon *a* and *b* are mutually exclusive. Or, if $f_{1a}' < f_{2a}'$ and $f_{1b}' < f_{2b}'$, they are skipped together.

(E1 and E3) are constitutive exons. Accordingly, each exon trio in TXdb was represented by two sequences (Figure 1D): an inclusion isoform (f_1) with all three exons; and a skipping isoform (f_2) comprising the flanking exons only. The first and last exons from every transcript were filtered out, because transcriptomic variability in these areas is primarily due to alternative transcription initiation or polyadenylation, rather than to AS per se. By using TXdb as a mapping database, we can approach every exon as an independent case to estimate its AS level.

Based on this concept, we adapted TXdb for detecting other types of AS (AA: alternative 3' splice site, AD: alternative 5' splice site. IR: intron retention). To analyze AD (Figure 1E) and AA (Figure 1F) we compiled exon duos (rather than trios), setting f_1 as the extended isoform (spliced via the proximal splice site), and f_2 as the shortened isoform (spliced via the distal splice site). In addition, to account for IR (Figure 1G), we defined f_1 as the intron-retaining isoform, and f_2 as the spliced isoform.

To estimate the extent to which the trio/duo assemblies can capture AS variability, we downloaded the compendium of AS events from the AStalavista website (http://genome.crg.es/astalavista/) (7). Using the RefSeq (hg18) AStalavista database, we counted the numbers of events that could be represented in the format of exon trio/duo. 74.14% of the data in AStalavista accounted for single-exon AS events, all of which were covered by TXdb with a single exon trio/duo. 9.82% of the events could be described by combining two entries in TXdb (e.g., consecutive CAs), and 2.9% corresponded to assemblies of three or more exon trios/duos. The rest (13.14%) corresponded to more complex AS events that could not be handled by TXdb.

To characterize complex AS events represented by combinations of multiple exon trios/duos, or overlapping AS events at the same locus, post-analysis would be necessary. For example, Figure 1H illustrates two consecutive exons that are alternatively spliced. If either of them is skipped, the respective exon trios would not pass the coverage cutoff (see Section 2.3 for details). However, if annotations exist, the inclusion ratio may be quantified based on substitute exon trios available in TXdb (Figure S1). By comparing the inclusion ratios of both exons, one may detect if they are mutually exclusive or

skipped together. It is important to note that the ability of SpliceTrap to detect complex AS events is limited, and depends on the availability of AS annotations to generate several trios/duos for each examined exon. For this reason, we recognize that some AS events may be overlooked, especially if they involve more than two consecutive alternative exons.

The final assembly of TXdb for hg18 comprises a total of 167,445 cassette-exon (CA) candidates, of which 11,812 have CA annotation, and the remaining 155,633 are annotated as constitutive exons (CS, to be examined whether they are in fact skipped). In addition, TXdb comprises 8,667 AA, 4,838 AD, and 1,170 IR candidates, based on annotations from dbCASE or RefSeq. All together, SpliceTrap contains 224,995 exon trios (or duos) embodying transcript variability from 182,560 human exons (Supplementary Table 1).

Finally, we wish to bring to the reader's attention that since ~6% of the exons in TXdb are uniquely annotated in dbCASE, a slight bias towards the 3' end of the transcript may exist, especially for AAs, ADs, and IRs, which are generally unique to dbCASE (Supplementary Table 1). TXdb is available on-line as part of the SpliceTrap package at http://rulai.cshl.edu/splicetrap/.

A Bayesian model to estimate inclusion ratios

In a paired-end RNA-seq experiment, a fragment is defined as a sequence segment encompassed between the first and last nucleotides of a read-pair. We assume that for each exon trio/duo, the positions of the mapped fragments follow a uniform distribution, and that their sizes follow a nearly normal distribution that depends upon the experimental protocol. Based on these assumptions, a fragment *j* can be described as a vector r_j: (b_j, s_j), where b_j and s_j denote the beginning position and size of the fragment, respectively.

Then, for every exon trio (or exon duo), we define the set of all possible isoforms as $= \{f_1, f_2\}$, where f_1 is an inclusion (or extended) isoform, and f_2 is a skipping (or shortened) isoform (Figure 1). The lengths and the relative expression levels of these

isoforms are $L = \{L_1, L_2\}$ and $E = \{e_1, e_2\}$. Accordingly, the probability of observing an isoform *i*, given the expression level *E*, can be written as:

$$P(f_i \mid E) = \frac{e_i \cdot L_i}{e_1 L_1 + e_2 L_2} \tag{1}$$

Let *m* be the number of fragments $R = \{r_j, j = 1, 2, \cdots, m\}$ that can be mapped to *F*. Given that for each fragment r_j: (b_j, s_j), b_j and s_j are independent, the probability of observing r_j, given an isoform f_i, is:

$$P(r_j \mid f_i, E) = P(b_j \mid f_i, E)P(s_j \mid f_i, E) = P(b_j \mid f_i, E)P(s_j) = \frac{1}{l_i} P(s_j) \tag{2}$$

where l_i is the effective length of f_i ($l_i = L_i - s_j + 1$), and $P(s_j)$ is the probability of observing a fragment size s_j in the experiment. Note that if only one end can be mapped to f_i, then $P(s_j)$ is set as 1 to ignore fragment-size information.

For all isoforms in *F*, we can write $P(r_j \mid E)$ as:

$$P(r_j \mid E) = \sum_{f_i \in F} P(r_j \mid f_i, E)P(f_i \mid E) = \sum_{f_i \in F} \left(\frac{1}{l_i} \cdot \mathrm{P}(s_j) \cdot \frac{e_i \cdot L_i}{e_1 L_1 + e_2 L_2}\right) \tag{3}$$

So for the whole data, we can write:

$$P(R \mid E) = \prod_{r_j \in R} \mathrm{P}(r_j \mid E) \tag{4}$$

Provided the prior distribution of E (see Section 2.4), a Bayesian posterior function can be written as:

$$P(E \mid R) \propto \prod_{r_j \in R} \mathrm{P}(r_j \mid E) \times P(E) \tag{5}$$

Then, we can maximize $P(E \mid R)$ to estimate the inclusion ratio e_1 for every exon.

Note that throughout the text we refer to $P(s_j)$ as FSM (Fragment-Size distribution Model), and to $P(E)$ as IRM (Inclusion-Ratio distribution Model), both of which are prior distributions and will be further described in Section 2.4.

Pipeline design

We designed a simple pipeline to run SpliceTrap (Figure 2). We started by mapping the read-pairs onto TXdb. For this purpose, we independently aligned every read to the inclusion/skipping isoforms in TXdb using Bowtie (29). Then, the fragments unambiguously mapped to single exons were used to build a FSM (Section 2.4).

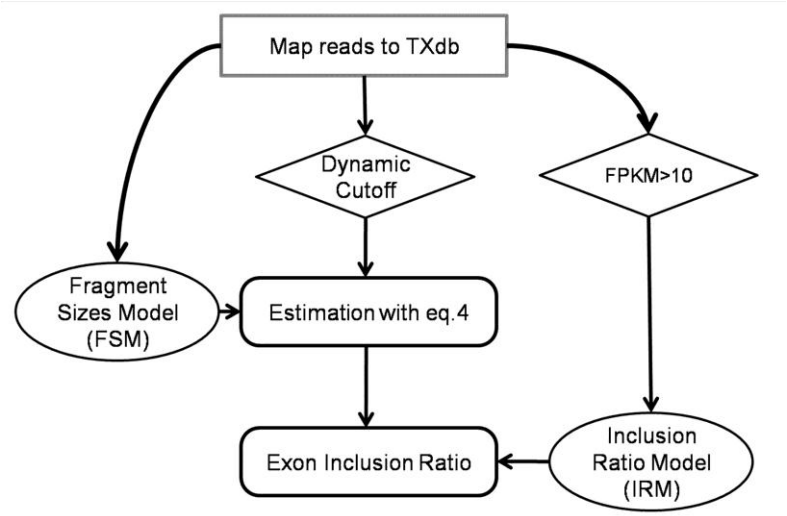


Figure 2. SpliceTrap Pipeline. This chart illustrates the order and interrelation among the different tasks performed by SpliceTrap. Squares represent mapping steps; diamonds are filtering steps; ellipses are prior-information models; rounded-corner rectangles represent steps in the Bayesian model.

To filter out poorly covered exon trios, we applied a dynamic, exon-size-dependent cutoff strategy (Supplementary Figure 2 and Section 3.1). Basically, we applied different coverage thresholds to every exon, such that the size of the exon and the coverage are inversely correlated. As an extreme example, an exon that is shorter than a read would need to be covered several times to be reliable. However, very long exons may be partially covered and still be reliable. Accordingly, we filtered exon trios with poorly covered flanking exons E1 and E3, but we did not require minimal coverage for exon E2 (i.e., the exon under consideration). This filtering method is intended to reduce noise resulting from rarely expressed transcripts, truncated transcripts, DNA contamination, wrongly mapped reads, etc., while avoiding unnecessary loss of information from exons with good coverage (details in Section 3.1).

Next, we maximized eq. 4 for every exon trio, utilizing all the mapped reads and the FSM to estimate the exon inclusion ratios. Finally, to reduce variability noise, we corrected the results with an inclusion-ratio distribution model (IRM) derived from high-confidence data (see Section 2.4).

Prior information models (FSM and IRM)

To generate FSMs, we took all the fragments uniquely mapped within the boundaries of constitutive exons (i.e., not spanning across splice junctions) and extracted the fragment sizes according to the positions of the reads. Finally, the occurrence of each fragment size was recorded to generate the distribution. The FSM distribution can be affected by variations in the experimental protocols. In previous studies, it was approximated as a normal distribution (100); however, to increase prediction accuracy, we chose to derive the FSMs directly from the dataset under study.

We generated IRMs for every type of splicing pattern separately (Supplementary Figure 3). Essentially, after mapping RNA-seq data onto TXdb, we selected the highest covered exon trios (FPKM>10) and estimated their inclusion ratios using eq. 4 (Supplementary Figure 3A-D). As a control, we also generated IRMs with Cufflinks (Supplementary Figure 3E-H). Notably, the distributions were very similar with both methods. To avoid overfitting, we smoothed the IRMs by fitting beta distributions (Supplementary Figure 3I-L) to the histograms, which were then used in subsequent correction steps. Note that there is no specific IRM for CS, because every CS is examined as a potential CA with a CA IRM.

Metrics for accuracy testing

To test the ability of SpliceTrap to discriminate alternative from constitutive exons with inclusion ratios, we designed a series of metrics based on TXdb annotations. The assumption is that exons annotated as CA are enriched within the fraction of exons with inclusion ratio $ir < 1$, and conversely, exons annotated as CS are included at approximately $ir = 1$.

Cassette exon discovery rate (CAD): This metric is analogous to the Positive Predictive Value (PPV). Given an $ir < 1$, all cassette exons above this ir are true positives (

CA_{ir} denotes the number), whereas all constitutive exons above the same ir are false positives (CS_{ir}); then,

$$CAD_{ir} = \frac{CA_{ir}}{CA_{ir} + CS_{ir}} \tag{6}$$

Constitutive exon discovery rate (CSD): By analogy to the False Positive Rate, above a certain ir ($ir < 1$), all constitutive exons are false positives (the number of which is denoted by $CS_{ir<1}$), whereas all constitutive exons at $ir = 1$ are true negatives ($CS_{ir=1}$), because these are reported as constitutively spliced, then CSD_{ir} can be written as:

$$CSD_{ir} = \frac{CS_{ir<1}}{CS_{ir<1} + CS_{ir=1}} \tag{7}$$

Specificity (SP): Using the definitions above, we calculate the specificity, which is nearly the converse case of CSD:

$$SP_{ir} = \frac{CS_{ir=1}}{CS_{ir<1} + CS_{ir=1}} \tag{8}$$

Results

SpliceTrap is a tool specifically designed to detect alternative splicing and quantify exon-inclusion ratios. Below, we present both simulations and data analysis demonstrating that SpliceTrap is highly accurate, reliable, and robust, and we also compare it to state-of-the-art RNAseq analysis tools.

Simulation of inclusion-ratio quantification

We carried out a simulation in order to test the accuracy of SpliceTrap compared to other methods. A series of exon trios was generated by analogy to TXdb. For every exon trio, the flanking exons (E1 and E3) were fixed to a size of 120 nt (the average exon size in TXdb) whereas the middle exons (E2) varied in size from 9 to 500 nt. For these isoforms, we set expression levels based on the distribution of inclusion ratios. We selected the IRM for cassette-exon events (CA), which is the most common AS type (Supplementary Figure 3I).

To simulate an RNA-seq experiment, we randomly fragmented the isoforms into overlapping fragments of sizes following a $N(200,15^2)$ distribution (by analogy to typical paired-end datasets), and preserved only the 75 (or 36) nt ends of each fragment as read-pairs. For every exon trio, the number of reads was adjusted to achieve exon coverage between 0 and 10. All together, we ran a total of 5555 simulations with different combinations of middle-exon size and coverage per tested method. For each simulation, 1000 repeats were made, and then the Pearson correlation coefficient (PCC) and the mean absolute error between the predicted and expected inclusion ratios were calculated for accuracy evaluation.

Table 1. Simulation Averages and Standard Deviations

Method	correlation coefficient		mean absolute error	
	36 nt	75 nt	36 nt	75 nt
RPKM	0.76 ±(0.18)	0.75 ±(0.14)	0.16 ±(0.11)	0.17 ±(0.06)
Cufflinks	0.83 ±(0.13)	0.78 ±(0.12)	0.11 ±(0.03)	0.16 ±(0.03)
Scripture	0.72 ±(0.22)	0.61 ±(0.19)	0.18 ±(0.10)	0.25 ±(0.09)
MLE	0.84 ±(0.14)	0.79 ±(0.12)	0.10 ±(0.05)	0.15 ±(0.03)
SpliceTrap	0.87 ±(0.14)	0.83 ±(0.12)	0.11 ±(0.05)	0.13 ±(0.04)

We evaluated five different methods (Table 1): A naïve method based on RPKM counts alone (87); Cufflinks (101) ; Scripture (42); a maximum likelihood estimation model (MLE) (SpliceTrap using uniform IRM and FSM models); and SpliceTrap. (see Supplementary Method 1.2 for the implementations of Cufflinks and Scripture). Our simulation demonstrated that SpliceTrap can outperform all the other methods, with higher PCCs and lower mean errors (Table 1). We observed that using 36nt or 75nt reads, the average PCC of SpliceTrap was the highest (0.83-0.87) compared to RPKM (0.75-0.76), Cufflinks (0.78-0.83) and Scripture (0.61-0.72). In addition, we noticed that by using prior information along with the MLE model (i.e., the full Bayesian model) we obtained better results compared to MLE alone (0.79-0.84) providing evidence for the contribution of the prior-information models to the estimations. A similar pattern can be found in the mean absolute errors, where SpliceTrap attained the lowest errors (0.11-0.13) compared to the rest of the tools (0.11-0.25).

Next, we carried out simulations for other types of AS patterns (Supplementary Table 3). We kept the same parameters, except for the IRM models, which were adjusted to each splicing type (Supplementary Figure 3). In all cases, the error means and PCCs obtained were similar to those calculated using a CA IRM with 36nt and 75nt reads. (Supplementary Table 3), indicating that SpliceTrap can be used to investigate different splicing patterns.

Notably, these simulations revealed a general association between the prediction accuracy, the size and the coverage of the exons (Supplementary Figure 4). Specifically, whereas smaller exons required higher coverages, low-coverage but larger exons achieved comparable accuracies. This resulted in a power-law-shaped surface, both for the mean error and the PCC, which was independent of the method used (Supplementary Figure 4). We took advantage of this observation and designed a dynamic-cutoff strategy accordingly. From the simulation results shown in Supplementary Figure 4A, curves were derived at different PCCs ranging from 0.1 to 0.9 (Supplementary Figure 2). In other words, the minimal coverage required to achieve corresponding PCCs for different exon sizes was recorded. For convenience, these dynamic cutoff curves are referred to as 0.1dc to 0.9dc throughout the text. Basically, for every exon in the data, we required a minimum coverage, depending on its size. Smaller exons required higher coverage, and larger exons required lower coverage. In short, this procedure should filter out most of the noise, and yet avoid unnecessary loss of exons that are partially covered, albeit by a sufficient number of reads.

Running SpliceTrap with RNA-seq data

To experimentally test SpliceTrap and compare it to other methods, we generated more than 60 million 36 nt paired-end reads using HeLa cell RNA (see Supplementary Methods 1.1). We applied SpliceTrap to these data, using dynamic cutoffs from 0.1dc to 0.9dc (Figures 3 and S2). We noticed that in general, the distributions of the inclusion ratios had a “U” shape (Figure 3A) which was also observed using Cufflinks (Supplementary Figure 3E-H) and in other studies based on ESTs (102). This means that in the sample analyzed, the exons tended to be highly included (i.e., constitutive) or fully

skipped from the transcripts. Nevertheless, a substantial proportion of the exons showed intermediate inclusion levels, regardless of the stringency of the dynamic cutoff.

In addition, we noticed that the number of selected exon trios did not vary dramatically within the range of lower dynamic cutoffs (0.1dc-0.6dc) although it dropped considerably above 0.7dc. This could be explained by the rapidly growing distances between the curves shown in Supplementary Figure 2. Based on these observations, we selected low (0.6dc), medium (0.7dc) and high (0.8dc) stringency dynamic cutoffs for further analysis.

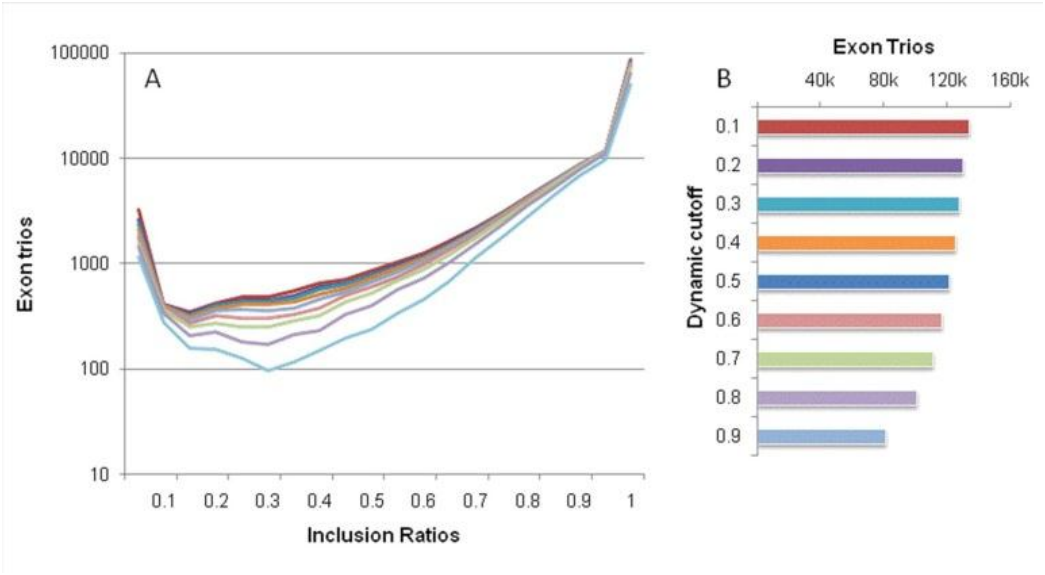


Figure 3. Distribution of inclusion ratios. (A) and number of detected exon trios (B) based on RNA-seq data from HeLa cells (36 nt paired-end). The colors correspond to dynamic cutoffs from 0.1dc to 0.9dc.

We ran Cufflinks and Scripture on the same datasets (provided with TXdb annotations), then we used three different cutoffs: FPKM=1, FPKM=2, FPKM=10 for Cufflinks; and RPKM=1, RPKM=2, RPKM=10 for Scripture. For every AS candidate, Cufflinks and Scripture reported the expression levels of the inclusion and skipping isoforms. We used these numbers to calculate inclusion ratios (See Supplementary Methods 1.2 for details).

Predicting known splicing patterns

We first tested the ability of SpliceTrap and other methods to detect known splicing events. In TXdb, every exon is assigned an annotation based on high-confidence ESTs and/or cDNAs (Supplementary Table 1). Our assumption is that exons annotated as cassette (CA) should be predominantly skipped (*ir<1*), whereas exons annotated as constitutive (CS) should be highly included at approximately *ir =1*.

Based on this premise, we extracted all the exons (CA and CS) and their inclusion ratios from the above results. CS exons were examined in SpliceTrap as potential CAs with a CA IRM. Therefore, in this assay, SpliceTrap was "blind" to the annotations in TXdb (CS or CA). To test the abilities of the different methods to discriminate between CA and CS, we calculated the CA discovery rate (CAD), CS discovery rate (CSD) and prediction specificity (SP) (see Section 2.5). Notably, for any selected threshold, SpliceTrap performed better at detecting low-included CAs, compared to the other tools (Figure 4A). Whereas Cufflinks and Scripture detected CAs in any *ir* range to a similar extent, SpliceTrap was more efficient at identifying known CAs as the *ir* decreased. For example, at *ir=0.5*, the SpliceTrap CAD value was ~0.6, and at *ir=0.1*, it was almost 1. On the other hand, SpliceTrap detected CSs almost exclusively at low inclusion ratios (Figure 4B), with CSD values below 0.1 at *ir=0.5* and ~0 at *ir=0.1*. Accordingly, SpliceTrap exhibited higher specificity than Cufflinks and Scripture (Figure 4C), achieving levels above 0.5 for *ir<0.5*.

In summary, SpliceTrap quantifications appear to be consistent with previous AS annotations; that is, most annotated CSs are included at around *ir=1*, whereas CAs are spread through the whole range of inclusion ratios (Figure 4). Using a U-shaped CA IRM (Supplementary Figure 3I) as prior information may have contributed to the prediction accuracy of SpliceTrap. Also, wrongly annotated CAs/CSs in TXdb or novel AS events might have affected the accuracy of the metrics.

Robustness and reliability of SpliceTrap

We evaluated the robustness of SpliceTrap estimations to technical variability among different replicates of the same experiment. To this end, we compared the results

obtained from two independent RNA-seq lanes from the same dataset (36 nt paired-end), generated under the same condition. The plots comparing both lanes, using SpliceTrap, Cufflinks, or Scripture with different thresholds, are shown in Figure 5.

Similar to Figure 4, SpliceTrap predicts most exons to be constitutively spliced (Figure 5A-C). Using the stringent cutoff, 78% of the exons were included at *ir≥0.9* in both experimental replicates. In contrast, only 30% of the exons for Cufflinks (Figure 5D-F) and 23% for Scripture (Figure 5G-I) showed *ir≥0.9* in both replicates. Correlations between inclusion ratios reported by different methods can be found in Supplementary Figure 8.

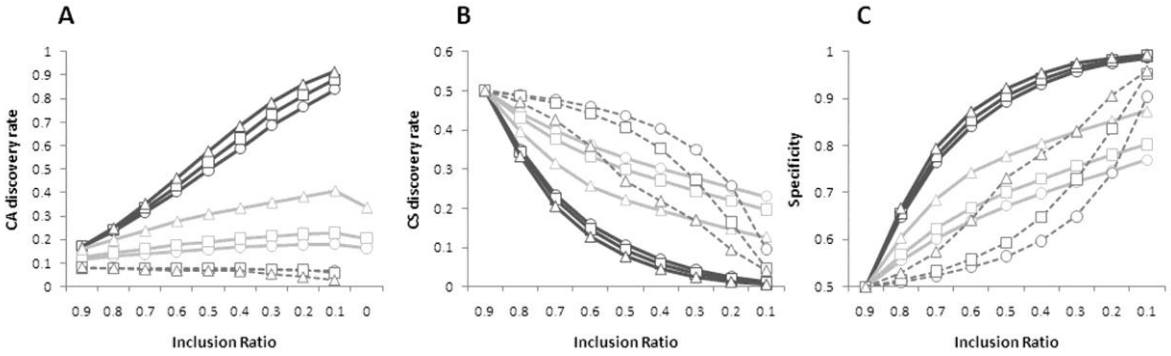


Figure 4. Predicting known splicing patterns using 36 nt paired-end reads from HeLa cells. (A) The cassette exon discovery rate, (B) constitutive exon discovery rate and (C) prediction specificity are shown as a function of the inclusion ratios (x axis) for SpliceTrap (black lines), Cufflinks (grey lines) and Scripture (dashed lines). Each method was applied using low (circles), mid (squares) and high (triangles) cutoffs. (0.6dc, 0.7dc, and 0.8dc for SpliceTrap, FPKM=1, FPKM=2, and FPKM=10 for Cufflinks; RPKM=1, RPKM=2, and RPKM=10 for Scripture).

Notably, SpliceTrap can reliably reproduce the results with increasing PCCs (0.74 to 0.77) depending on the threshold. In contrast, Cufflinks achieved a maximum PCC of 0.7 (Table 2), but only when using a high threshold (FPKM=10). This means that it can achieve levels similar to SpliceTrap at the expense of the number of results: whereas SpliceTrap reported 97,068 exons at PCC=0.74, Cufflinks reported only 11,606 exons at PCC=0.7. Scripture performed with high PCCs, although the predicted scores suggested that the majority of human exons are alternatively spliced (Figure 5), which is not in

agreement with previous transcript annotations. Note that the correlation values here concern the reproducibility, which is not neccesarily related to the accuracy presented in Table 1.

Of note, Cufflinks achieved a high PCC in reproducing the expression levels of the inclusion (0.91) and skipping (0.81) isoforms (Supplementary Figure 7), suggesting that Cufflinks has a higher robustness in detecting full transcript expression than AS.

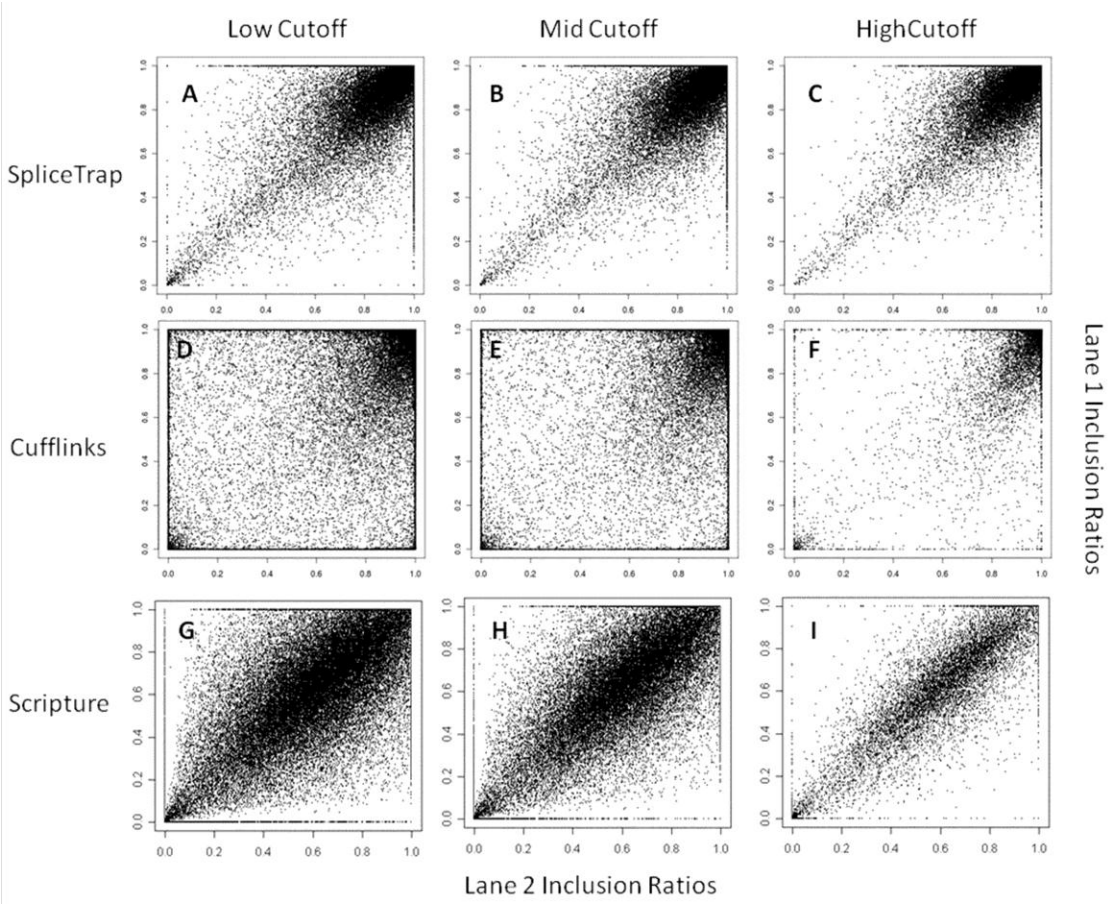


Figure 5. Robustness of the inclusion ratio estimations. The charts illustrate the correlations between the inclusion ratios calculated in two independent RNA-seq lanes (36 nt paired-end data from HeLa cells). (A-C) SpliceTrap at 0.6dc, 0.7dc, and 0.8dc, (D-F) Cufflinks at FPKM=1, FPKM=2 and FPKM=10, (G-I) Scripture at RPKM=1, RPKM=2 and RPKM=10.

Finally, we wanted to rule out dependencies between the net expression levels and the inclusion ratios detected by SpliceTrap and the other tools. To this end, we ranked all calculated inclusion ratios from one lane in Figure 5(A,D,G) according to the expression

levels of the full-length exon trios reported by Cufflinks. We smoothed the inclusion ratios with a sliding window of size 970 for SpliceTrap, 522 for Cufflinks, and 704 for Scripture, corresponding to the number of exons analyzed divided by 100 (see the exon numbers for the lowest cutoffs in Table 2).

We observed that the inclusion ratios calculated with SpliceTrap were independent of the expression levels, with a constant average rate of ~0.95 (Supplementary Figure 5). In contrast, the inclusion ratios calculated with Cufflinks decreased proportionally to the expression levels. The inclusion ratios calculated with Scripture were also constant; however, they averaged around 0.5, meaning that most exons in the data are viewed as alternatively spliced.

Table.2. Comparison of two replicates (36 nt paired-end reads)

Method	Cutoff	Exons	PCC
Cufflinks	FPKM=1	52243	0.41
Cufflinks	FPKM=2	38140	0.49
Cufflinks	FPKM=10	11606	0.7
SpliceTrap	0.6dc	97068	0.74
SpliceTrap	0.7dc	90896	0.75
SpliceTrap	0.8dc	80052	0.77
Scripture	RPKM=1	70466	0.83
Scripture	RPKM=2	49816	0.87
Scripture	RPKM=10	14022	0.92

In conclusion, SpliceTrap can detect AS events in a reliable and reproducible way, compared to Cufflinks and Scripture, which can be adapted to identify AS events, albeit with lower accuracy.

Discussion

SpliceTrap is a computational tool fully dedicated to mapping AS activity based on paired-end RNA-seq data. Unlike other available tools, SpliceTrap can quantify splicing ratios under single cellular conditions, because it does not require a background. Rather than reporting background-based read densities, SpliceTrap utilizes Bayesian statistics to summarize inclusion probabilities derived from every single read-pair. For this reason, SpliceTrap is also insensitive to transcript expression levels.

SpliceTrap was specifically designed to accurately quantify alternative splicing at the single exon level. To achieve this goal, we started by describing the problem with a statistical model based on exon trios/duos, instead of full transcripts. To reduce the number of false positives and yet minimize the loss of information, we applied dynamic cutoffs derived from simulation, rather than using fixed cutoffs. Finally, we adjusted the results using specific inclusion-ratio models for different AS patterns. In theory, full-transcript tools like Cufflinks and Scripture can also be adapted to calculate inclusion ratios with TXdb annotations (Supplementary Method 1.2), However, these tools were originally designed and optimized for transcript-level estimation, and our analysis indicates that they are less accurate than SpliceTrap for the specific problem of calculating inclusion ratios.

We have shown that it is possible to approach every exon as a separate problem, and yet quantify its inclusion ratio without knowledge of the full transcript structure. Given that our quantitative units are the exons, we can disregard this information, though it is certainly important for transcript-level quantification.

SpliceTrap can detect different splicing patterns. Even though the algorithm was developed to detect single cassette exons, we could adapt it to other types of AS, such as alternative 3’/5’ splice sites and intron retention. In combination, these patterns account for ~75% of all known AS in eukaryotes (Sammeth, et al., 2008). Additionally, some of the complex AS patterns involving multiple exon trios/duos are also detectable(Supplementary Figure 6). However, this ability is limited and depends on the availability of TXdb annotations. To resolve complex or overlapping AS events, it is

necessary to combine and compare inclusion ratios from different exon trios/duos. And one may need to focus on exon trios/duos supported by junction reads to improve the detection specificity. These post-analysis steps can filter out possible bias and differentiate true events. However, this process is case-specific and may not always give a solution. Also, some of the complex events cannot be handled by exon trios/duos, as mentioned in Section 2.1.

At this stage, SpliceTrap does not offer an option for *ab initio* exon prediction. We chose to focus on a set of ~200,000 well characterized exons, and designed a transcript database (TXdb) as a mapping reference. In this way, we sought to reduce ambiguities generated by rarely expressed isoforms, especially during the mapping procedure. Because TXdb is a collection of exon trios/duos, in the future it can be expanded by adding newly discovered or predicted splice juctions, such as those derived by *de novo* mapping software like TopHat, or novel exons predicted by gene-prediction tools, e.g., GENSCAN (103). In this way, the depth and sensitivity of SpliceTrap can be enhanced.

SpliceTrap is based on the assumption that the reads are uniformly distributed within the exon trio/duo. Although the uniformity in a small region is presumably a better assumption than in a full transcript, this factor will still bias the results and should be considered in future versions of the model.

SpliceTrap can be implemented on-line through the CSH Galaxy server (http://cancan.cshl.edu/splicetrap) or downloaded at http://rulai.cshl.edu/splicetrap/. Both versions are easy to operate and require a small number of input parameters, considerably reducing the setup time. The running time of SpliceTrap is about 3 hours for one lane of reads (20×2 million 36 nt paired-end reads), and approximately 12 hours for three lanes (60×2 million 36 nt paired-end reads). These measurements were taken on a single AMD CPU of 2GHz, with 8GB usable memory. Less than 500MB memory was used during the runs, although large hard-disk space was needed (10GB to 30GB). The running time is approximately linearly dependent on the number of mapped reads and database size, but

it is not affected by the read size. Also, SUN Grid engine(SGE) qsub is supported for parallel computing.

Acknowledgment

We thank Assaf Gordon for providing access and support to the CSH Galaxy server and Chaolin Zhang for providing dbCASE. Thanks also to Chaolin Zhang and Justin Kinney for helpful discussions.

Funding

This work was supported by the National Institutes of Health [GM74688 to M.Q.Z.].

Supplementary Methods

RNA-Seq dataset preparation

To evaluate the ability of SpliceTrap to detect known alternative splicing events, we generated a dataset of more than 60 million mRNA paired-end reads of length 36 nt per end (3 lanes), using Illumina GAIIx paired-end sequencers (Supplementary Table 2).

We obtained cytoplasmic mRNA from HeLa cells by gentle fractionation in lysis buffer (10 mM HEPES pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.5 % (v/v) NP-40). We pelleted the nuclei at 2300 g for 5 min and extracted cytoplasmic RNA from the supernatant using Trizol, followed by treatment with DNase I (Promega). This procedure enriches for cytoplasmatic mRNA, reducing the amount of nuclear splicing precursors and intermediates.

We used 4 μg of cytoplasmic RNA from each sample to prepare paired-end mRNA-seq libraries, following the manufacturer's instructions (Illumina). Briefly

For each sample, cDNA was prepared and sequencing adapters were ligated. The DNA was then separated on a 2% agarose TBE gel, and a band around 250 nt (corresponding to a fragment size of 185 nt plus adaptors) was excised and subjected to 15 cycles of PCR amplification to increase the amount of material.

The amplified DNA was then added to an Illumina flow cell and cluster generation was carried out on an Illumina station. Flow cells were then subjected to paired-end sequencing on an Illumina GA IIx sequencer with chemistry version 2 or 3.

The datasets are available at http://rulai.cshl.edu/splicetrap/.

Estimating inclusion ratios with Cufflinks and Scripture

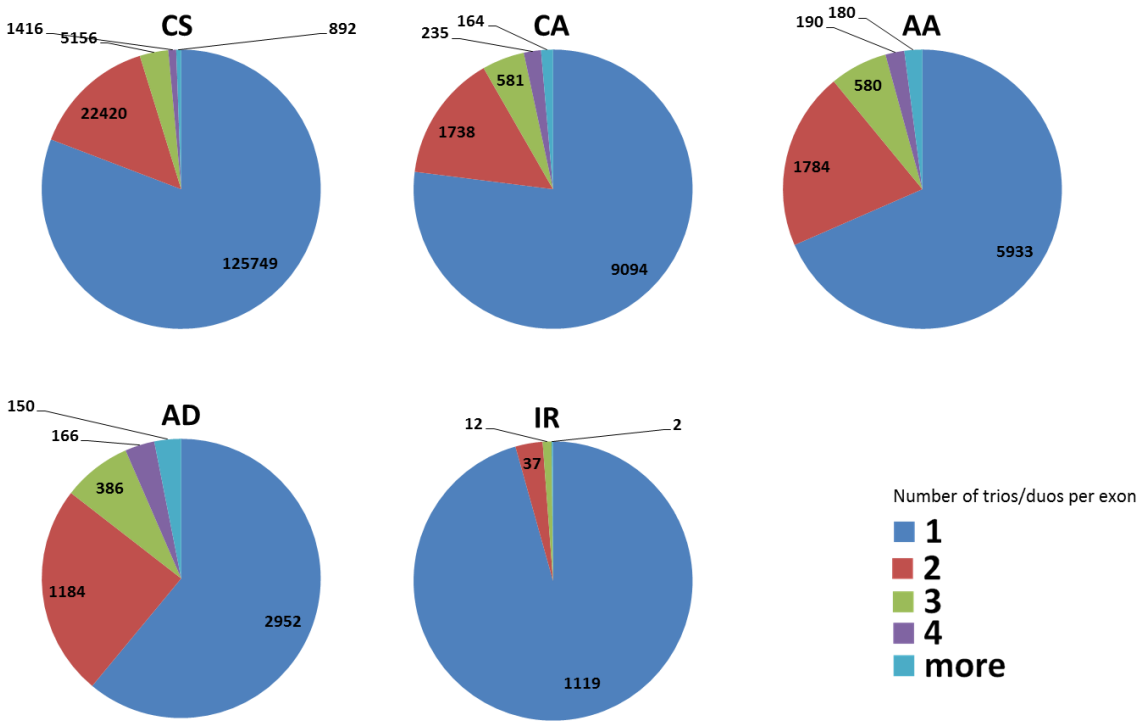
For RNA-seq data, we first mapped the reads with TopHat (32) providing the known junction database derived from TXdb. The expected mean inner distance between mate pairs was set as 100nt, which is the property of our data, and all the other parameters were kept as default. After that, we ran either Cufflinks or Scripture on the accepted hits.

In both software, TXdb annotations were provided in GTF format for quantification purposes, and default parameters were used in the running. Because these tools can report read-density values for inclusion (e_I) and skipping (e_S) isoforms separately, we can calculate the inclusion ratios (ir) with its definition:

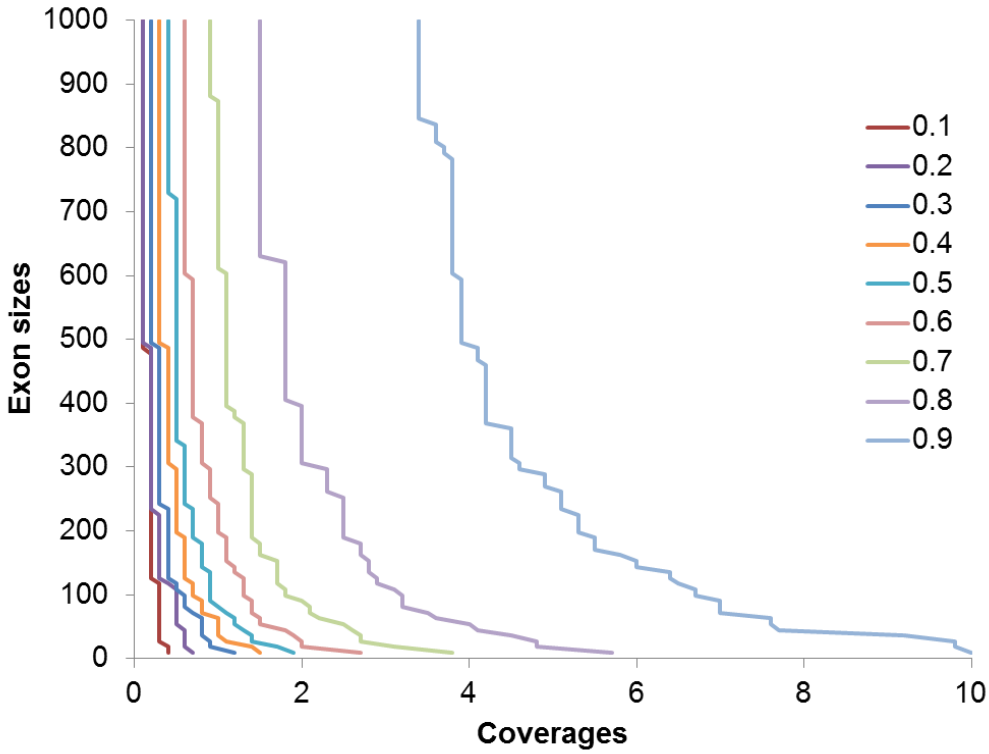
$$ir = \frac{e_I}{e_I + e_S}$$

In the simulation, TopHat was not used because the mapping results were produced by the simulations. Simulated gene structures were provided to Cufflinks and Scripture as well.

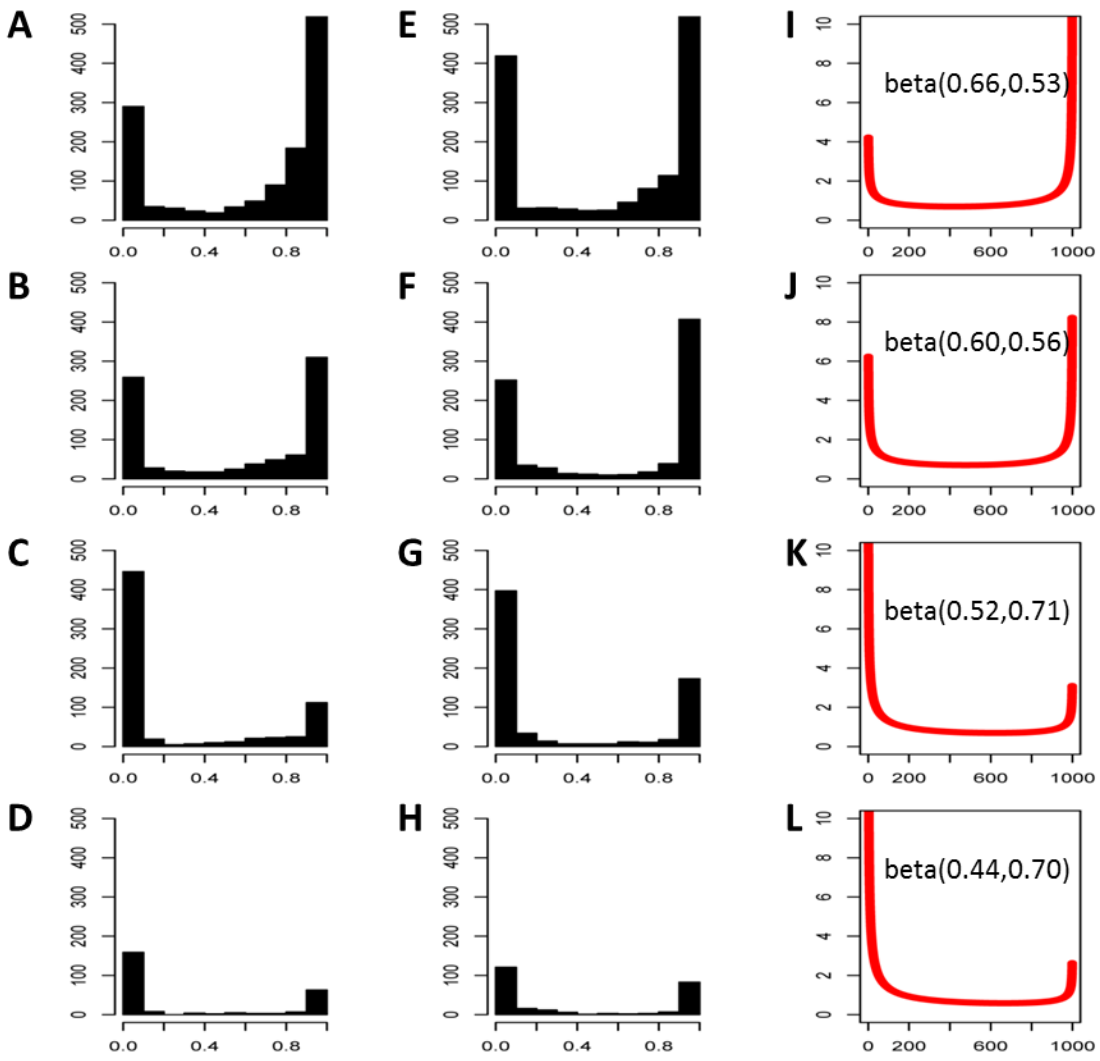
Supplementary figures



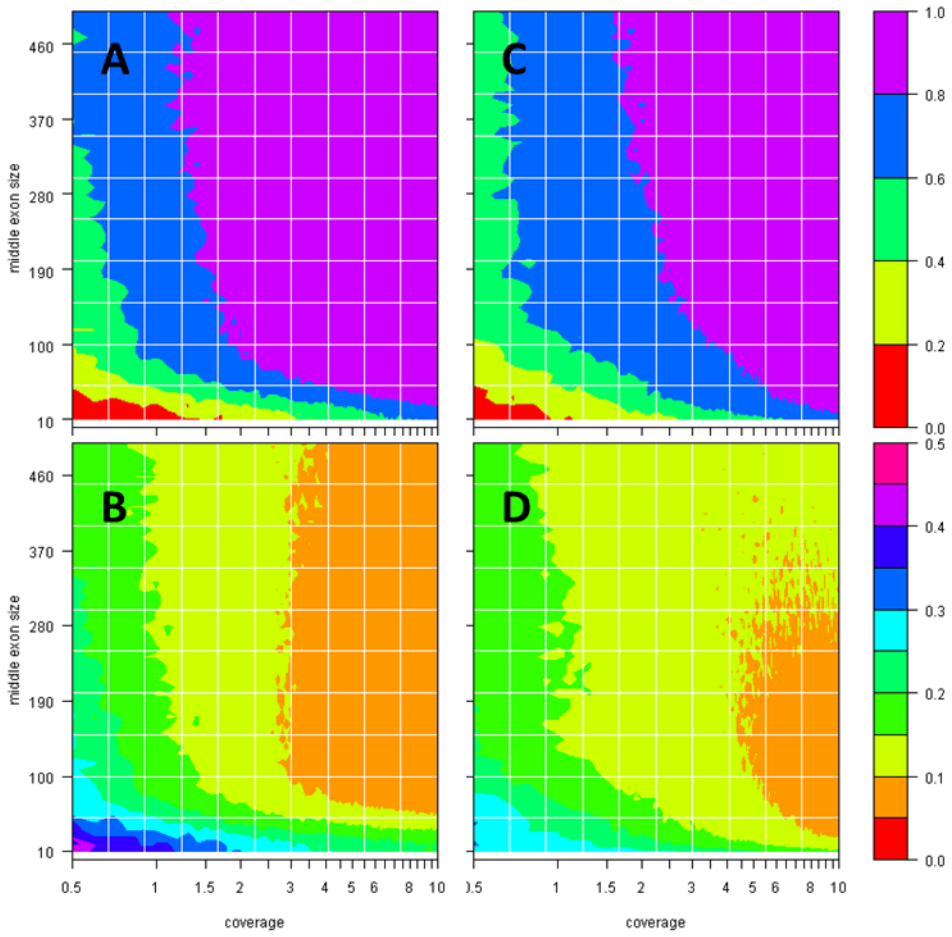
Supplementary Figure 1. Number of trio or duo assemblies per exon in TXdb . Absolute values are shown in the chart for exons annotated as constitutive (CS), cassette (CA), alternative 3' splice sites (AA), alternative 5' splice sites (AD) and intron retention (IR).



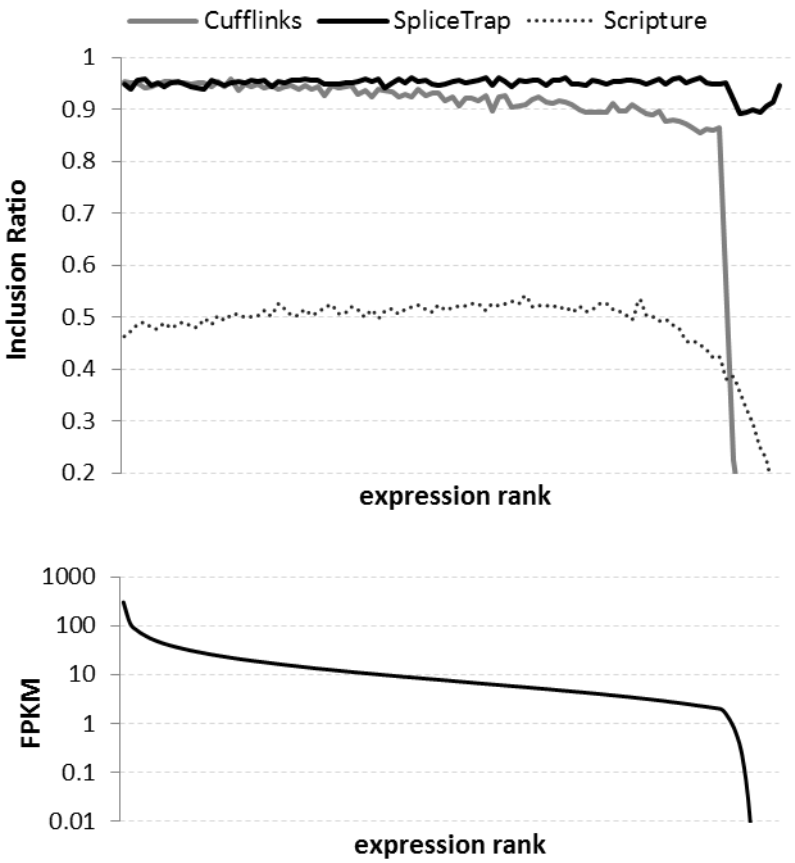
Supplementary Figure 2. Dynamic cutoff curves. The curves were generated according to Pearson correlation coefficients (PCC) ranging from 0.1 to 0.9 in Supplementary Figure 4A. The minimal coverage required to achieve corresponding PCCs for different exon sizes is recorded. Throughout the text, these cutoff curves are referred to as 0.1dc to 0.9dc.



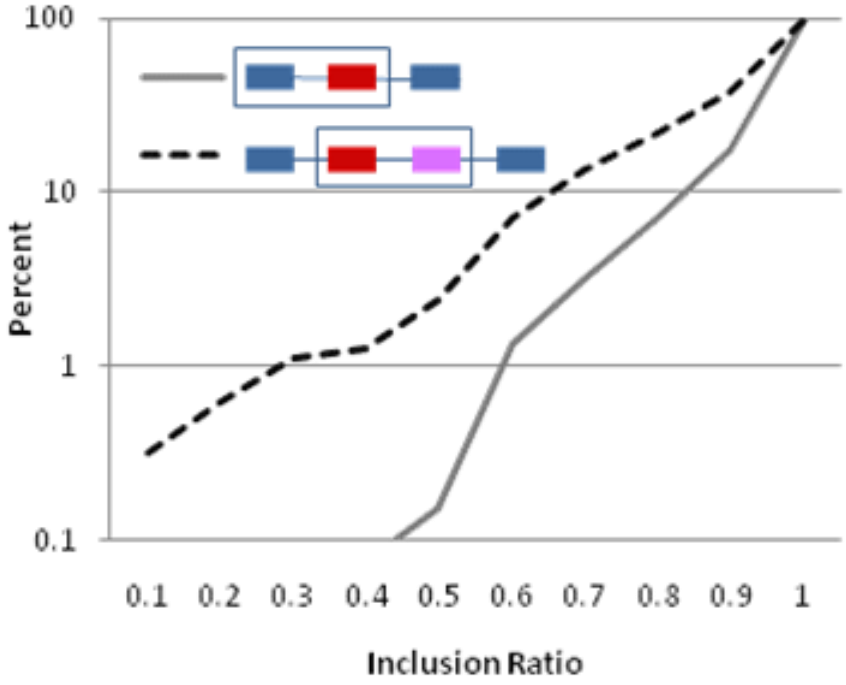
Supplementary Figure 3. IRM models from RNA-Seq data (36 nt paired-end reads).
Inclusion-ratio histograms for exon isoforms with FPKM>10 are illustrated for (A-D) SpliceTrap and (E-H) Cufflinks. Beta distributions were fitted to histograms A-D and used as IRMs (I-L); for every curve, the beta distribution parameters are shown. In the illustration, A,E,I represent cassette exons; B,F,J alternative 3' splice sites; C,G,K alternative 5' splice sites; and D,H,L intron retention events.



Supplementary Figure 4. Estimation accuracy is related to both coverage and exon size in the simulations. The contour images represent the correlation coefficients (A,C) and mean absolute errors (B,D) for (A,B) SpliceTrap and (C,D) Cufflinks. Y axes denote the exon sizes, and X axes represent the coverage. The color bars represent either Pearson correlation coefficients (upper) or mean absolute errors (lower). All the panels are from the simulation with 36 nt paired-end reads.

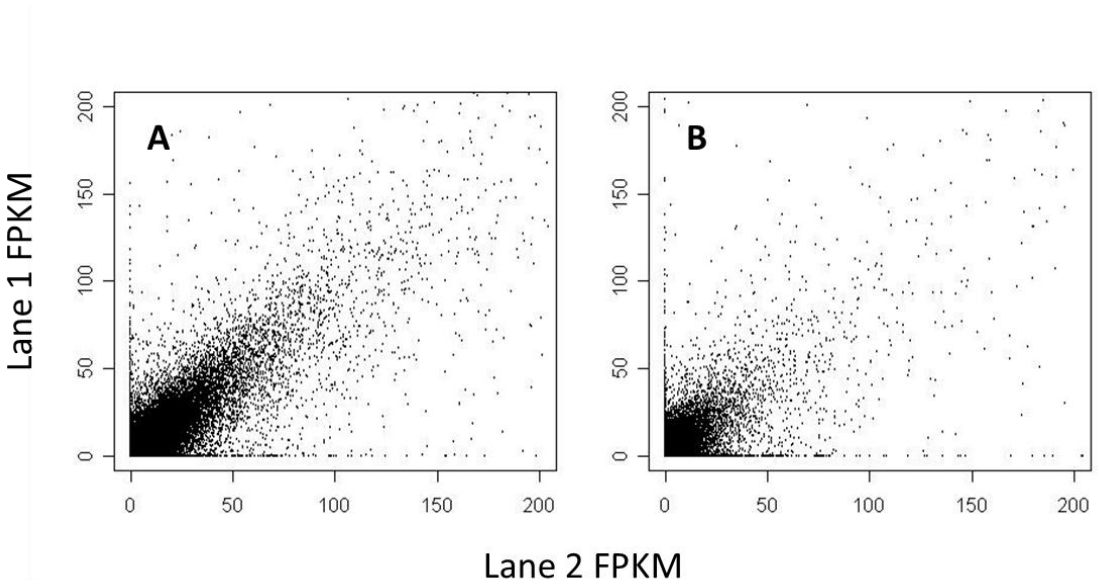


Supplementary Figure 5. Correlation between inclusion ratios and transcript expression. Expression levels were shown in upper panel: SpliceTrap (black), Cufflinks (grey) and Scripture (dashed). We ranked all calculated inclusion ratios from one lane (36 nt paired-end reads) in Figure 5(A,D,G) according to the expression levels of the full-length exon trios reported by Cufflinks (FPKM values shown in lower panel). 100 windows were used to smooth the curves.



Supplementary Figure 6. Correlation of inclusion ratios among double cassette exons (DCA, dashed line) and constitutive/cassette pairs (CSCA, full line).

SpliceTrap was used to calculate inclusion ratios on the 36 nt paired-end RNA-Seq dataset from HeLa cells (Supplementary Method 1.1). Subsequently, 2603 DCAs and 27557 CSCA pairs were extracted from TXdb. After analyzing the data only 637 DCAs and 5266 CSCAs passed the dynamic cutoff 0.6dc. We defined that if inclusion ratios of both exons in an exon pair are lower than a certain threshold value (x-axis), then this exon pair is considered as co-skipped. The percentages of co-skipped exon pairs (y-axis) were then recorded for different thresholds. For example, 5% of DCAs were co-skipped at a 0.5 inclusion ratio or less (31 cases), whereas only 0.15% of CSCA pairs presented this behavior.



Supplementary Figure 7. Reproducibility of transcript expression level results using Cufflinks. (A) the long isoform, or (B) the short isoform. X and Y axes stand for FPKM values in two independent lanes of data (36 nt paired-end RNAseq data from HeLa cells; see Supplementary Method 1.1 for details). The PCCs are 0.91(A) and 0.81(B).

Supplementary Table 1. TXdb composition

Type of splicing event	Refseq only	dbCASE only	both	total
CS	155633	0	0	155633
CA	2213	3835	5764	11812
AA	868	4180	3619	8667
AD	591	2419	1828	4838
IR	0	1170	0	1170
total	159305	11604	11211	182560

CS:constitutive exons
CA:cassette exons
AA:alternative 3' splice sites
AD:alternative 5' splice sites
IR:intron retention

Supplementary Table 2. Quality of the RNA-seq data

Total reads	60025808×2
Low quality	11230674
ambiguous	38213976
Paired	28798084
Unpaired	16338575
Used in estimation	45136659

Supplementary Table 3. Pearson correlation coefficients for simulations done using 36nt (A) and 75nt (B) paired-end reads, and absolute mean errors for different AS event types with 36nt (C) and 75nt (D) paired-end reads. Simulation results for CA are in the main text.

(A)

Method	AD	AA	IR
RPKM	0.76±(0.18)	0.75±(0.16)	0.78±(0.17)
Cufflinks	0.84±(0.13)	0.82±(0.15)	0.85±(0.12)
Scripture	0.72±(0.22)	0.72±(0.22)	0.76±(0.21)
SpliceTrap	0.88±(0.13)	0.86±(0.16)	0.89±(0.12)

(B)

Method	AD	AA	IR
RPKM	0.77±(0.14)	0.76±(0.14)	0.78±(0.14)
Cufflinks	0.79±(0.11)	0.78±(0.11)	0.80±(0.10)
Scripture	0.66±(0.18)	0.63±(0.19)	0.68±(0.18)
SpliceTrap	0.85±(0.12)	0.84±(0.12)	0.86±(0.11)

(C)

Method	AD	AA	IR
RPKM	0.16±(0.12)	0.16±(0.13)	0.16±(0.12)
Cufflinks	0.11±(0.03)	0.12±(0.04)	0.11±(0.03)
Scripture	0.18±(0.10)	0.18±(0.10)	0.16±(0.07)
SpliceTrap	0.10±(0.05)	0.11±(0.06)	0.10±(0.04)

(D)

Method	AD	AA	IR
RPKM	0.18±(0.07)	0.17±(0.06)	0.17±(0.07)
Cufflinks	0.17±(0.03)	0.16±(0.03)	0.16±(0.03)
Scripture	0.24±(0.07)	0.25±(0.08)	0.23±(0.07)
SpliceTrap	0.12±(0.04)	0.13±(0.04)	0.12±(0.04)

CA:cassette exons
AA:alternative 3' splice sites
AD:alternative 5' splice sites
IR:intron retention

Chapter 5 :

Systematically Discovery of Conserved Alternative Spliced Exons in the Mammalian Genome

So far, a pipeline to detect and quantify AS events from RNA-Seq data has been established. It has three major components: (1) A fast and sensitive splice mapping algorithm: OLego, which is able to use very small seeds to achieve high sensitivity on detecting exon junctions and small exons; (2) A set of in-house scripts to discover novel exons and novel splicing events from alignment results; and (3), SpliceTrap, an accurate method to estimate exon inclusion ratio from the data. This pipeline is a combination of splicing-centric tools. In this Chapter, the pipeline will be applied on a public RNA-Seq dataset for a systematical study of alternative splicing in mammalian genome.

Abstract

Alternative spliced (AS) exons in mammalian transcripts can either increase the diversity of protein products, or lead to non-sense mediated mRNA decay (NMD) of the transcripts. Recent technological advances in high-throughput mRNA sequencing (mRNA-Seq) give us an opportunity to investigate the mammalian transcriptome in an unprecedented depth and resolution, leading to the discovery of many novel exons, including low abundance

AS exons frequently subject to NMD. It is still unclear how many exons remain to be discovered and importantly, how many of those are likely to be functional. To address these questions, we aligned 3.7 billion mRNA-Seq reads from the BodyMap 2.0 data sequenced from 16 human tissues, using OLego, a program we developed recently to improve the sensitivity and accuracy in mapping exon-junction reads. This is followed by a pipeline that allows us to discover novel alternative exons, evaluate evolutionary selection pressure, and detect NMD isoforms. As a result, we identified 120,110 cassette exons in the human genome, including 45,687 exons without previous evidence of inclusion and 36,841 exons without previous evidence of exclusion. Using a mixture model to characterize sequence conservation across vertebrate species, we estimate that ~16% of these cassette exons are under significant purifying selection pressure. Strikingly, we identified over 3,000 cassette micro-exons smaller than 30 nt, including 105 exons with a length of 6 nt. Because of the minimal information that can be possibly encoded in this set of exons, they serve as an excellent model to study their functional significance and mechanism of AS regulation.

Introduction

For a majority of alternative splicing events, especially those newly identified in deep sequencing, the minor isoform is rare (104), frame-shifting, unproductive in terms of protein coding, and frequently coupled with NMD (105-108). These exons are predicted to be weakly deleterious, and a majority of them are expected to be eliminated during evolution (109).

In Drosophila, NMD is critical for sex determination (110,111),. In mammals, NMD is important to clear aberrant splicing products, but the global picture regarding the function of regulated AS events coupled with NMD is unclear (112). For a majority of NMD-coupled AS events, the NMD-isoform is low independent of the action of NMD (113), which is consistent with the idea that they are not regulated. An emerging idea is that while alternative splicing provides evolutionary plasticity, aberrant splicing could also cause various genetic diseases and be a burden of health.

In general it is difficult to distinguish regulated AS events from splicing errors. One measure is the conservation of splicing pattern between different species. However, as sequencing depth increases, it is more and more likely to detect the low abundant isoforms in both species by chance (107,114), and does not provide strong evidence for function.

Individual examples of regulated AS events coupled with NMD have been identified, and they overlap with very conserved exons and flanking intronic sequences, similar to regulated AS exons that produce alternative protein products (79,115,116). Most prominent examples are those in splicing factors, and in particular SR proteins and hnRNP proteins, which are important to maintain homeostasis of these proteins (106,117).

In this study, we mapped 3.7 billion of RNA-Seq reads from BodyMap 2.0 data with OLego, and used a pipeline to discover novel alternative exons, evaluate evolutionary selection pressure, and detect NMD isoforms. As a result, we identified more than 80,000 novel cassette exons. We then employed a Gaussian mixture model to characterize sequence conservation across vertebrate species. We estimate that ~20% of these cassette exons are under significant purifying selection pressure. We also identified over 3,000 cassette micro-exons smaller than 30 nt, which can serve as an excellent model to study their functional significance and mechanism of AS regulation.

Methods and Materials

BodyMap 2.0 data

The BodyMap 2.0 RNA-Seq data were sequenced from 16 human tissues (adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells) in 2010 by Illumina (ArrayExpress ID: E-MTAB-513). 1,278,683,163 reads were sequenced with 50 nt paired end tags (2x50 reads) and 1,263,636,284 reads were sequenced with 75 nt single end tags (1x75 reads). In other words, there were ~160 million reads sequenced for each tissue. In addition, 16 lanes of stranded 100 nt single end data (1x100 reads) were sequenced for 16-tissue

mixtures (1,194,539,556 reads). The numbers for each sample can be found in Supplementary Table 1. All the reads were used in this study.

Mapping the data with OLego

We first mapped the reads to the reference genome (hg19) using OLego (v1.1.1), a fast and sensitive tool to map spliced reads (39). To maximize the sensitivity, a comprehensive exon junction database was provided to OLego. The database was collected from mouse, human, and rat according to alignment of RefSeq transcripts, mRNAs and ESTs (79). Totally 356,777 unique exon junctions were included.

To take advantage of OLego's ability to detect ultra-small exons, we used option "–e 6" to allow identification of exons as short as 6 nt. For stranded libraries, "--strand-mode 1" was used. For paired end data, each end was mapped separately. Then both ends were merged according to their distance and orientations to eliminate ambiguities.

Identification of exons and AS events

Unique junction reads were extracted from the SAM format(67) alignment output, and all junction reads from 36 experiments were merged for a maximum set of novel junctions. Afterwards, these junctions, together with known junctions and alignment of known transcripts from Refseq, mRNAs and ESTs, were used to identify alternative splicing events using a similar splicing graph-based approach described previously (79). Specifically, to detect novel alternative splicing events, we used information from mRNA-Seq data. First novel exons were predicted by pairs of 3' and 5' splice sites separated by ≤ 300 nt. Previously defined exons and introns were combined with novel exon and introns to generate splice graphs and detecting alternative splicing events. Different from detection of AS events from ESTs and mRNAs, mRNA-seq reads in general do not span whole exons or multiple junctions. Therefore, we define novel AS events if each exon or intron involved in the AS is supported by either mRNA/ESTs or mRNA-seq reads. For the AS events analyzed in this study, we require each intron to be supported by ≥ 1 mRNA/ESTs, or two mRNA-seq junction reads, or both. Cassette exons, tandem cassette exons, alternative 5' and 3' splice sites, mutually exclusive exons, and retained introns were extracted from the data, and only cassette exons were used in

the next steps. Compared to transcript based reconstruction methods, this local approach can identify exons and AS events with higher sensitivity.

Estimation of the inclusion ratios

TXdb was constructed with all cassette exons identified from the data using the TXdbgen script provided in the SpliceTrap (v0.90.5) package (10). Both inclusion and exclusion isoforms for every exons were stored in this database so that SpliceTrap can use the information to estimate inclusion ratios for all detected cassette exons in each of the 32 samples (the tissue specific samples). Default options were used for SpliceTrap.

Detection of AS events that trigger NMD

A transcript is predicted to be targeted by the NMD pathway if it harbors a premature stop codon (PTC) ≥50 nt from the last exon junction. Conceptually, we define that an alternative splicing event is able to "trigger" NMD if (i) a pair of transcripts that differ only in the AS region but the same in other regions; and (ii) the transcript supporting one isoform is subject to NMD, while the transcript supporting the other isoform is not.

AS coupled with NMD has been studied previously (106,114). Annotations of NMD in these studies are generally incomplete because they rely on sequenced transcripts (mRNAs and ESTs). However, the current mRNA/EST sequence database is highly incomplete in terms of full-length transcript, as suggested by many novel alternative splicing events defined in mRNA-Seq data (this study and (74)). In addition, the current methods require both isoforms observed in transcripts with complete ORF. However, the NMD isoform in general has low abundance. Furthermore, the current methods did not formally distinguish AS events that "trigger" NMD from those related to NMD transcripts. An NMD exon detected by these approaches might be actually triggered by an upstream exon.

Therefore, we implemented a pipeline to evaluate AS events and whether they can "trigger" NMD with two treatments important for high accuracy and sensitivity. (i) For each transcript supporting one isoform of the AS event, we generate the transcript supporting the other isoform *in silico*, and predict if the AS triggers the NMD of the

transcript. (ii) In current methods, PTCs are typically determined by searching the longest ORF. This can result in false negative predictions when NMD-triggering AS events produce downstream PTCs which are not far away from the start codon, because an alternative AUG can frequently be used to predict the longest ORF, and the NMD-triggering events can be mistakenly classified as affecting 5' UTR, which do not trigger NMD by definition. In general, the ribosome initiates translation at the 5' most AUG of an mRNA, but sometimes it can skip one or more AUGs (118). In this study, to determine the open reading frame (ORF) of translation, we start translation of each transcript from the first overlapping start codon according to RefSeq and UCSC transcripts, rather than using the longest ORF.

For each AS event, we then count

i) Number of transcripts that will be targeted by NMD when the exon is included, but not if the exon is excluded (N_{In}).
ii) Number of transcripts that will be targeted by NMD when the exon is excluded, but not if the exon is included (N_{Ex}).
iii) Number of transcripts that encodes protein product for both isoforms (N_c).

An AS event is defined as

i) NMD upon inclusion (NMD_in), if $N_{In} > 0$, $N_{Ex} = 0$, and $N_{In} > 2 \times N_c$,
ii) NMD upon exclusion (NMD_ex), if $N_{In} = 0$, $N_{Ex} > 0$, and $N_{Ex} > 2 \times N_c$.
iii) coding if $N_{in} = 0$, $N_{In} = 0$, and $N_c > 0$
iv) other, for all remaining cases.

Identification of AS events under purifying selection

We note that alternative splicing events that trigger NMD upon exon inclusion are not under selection pressure of protein coding. Therefore, alternative splicing of these exons are presumably functional if they are conserved between different species (i.e. human and mouse). We obtained 161 NMD_in cassette exons with conserved splicing patterns as a positive control dataset for regulated AS events. These exons are required to present in both human and mouse and have no overlap with CDS regions. To get a negative control

set, we get a set of 2,244 exons that are constitutively spliced in both human and mouse (≥10 supporting transcripts). For a better interpretation of the intronic conservation, we also require that their flanking ±200 nt intronic sequences have no overlap with other exons. The selection pressure of each exon is measured by the average 46-way vertebrate phyloP score in flanking intronic sequences (±200 nt). Since phyloP score measures selection pressure of individual nucleotides (in contrast to phastCons score)(73,119), it is possible to distinguish the selection pressure of different codon positions in exons.

We also inferred the wobble position of exons using two approaches. In the first approach, we directly inferred the open reading frame from supporting transcripts, and inferred the reading frame of the cassette exon (ORF inferred). In the second approach, we calculated the average phyloP score for the three reading frames, and inferred the frame with the smallest score as the wobble position (for most constitutive exons, this position can be very easily identified (phylogenetically inferred). These two approaches agree on 93.4% of exons with orthologous sequences in mouse and human (when the reading frame can be inferred directly from coding transcripts), but the phylogenetic approach is not limited to transcripts with known ORF, and is able to give a direct comparison between coding and noncoding exons. We therefore use the average phyloP score of the phylogenetically inferred wobble position as an approximation of selection pressure driven by splicing regulation. In cases where the two approaches do not agree, the phyloP score of the ORF inferred wobble position is typically very close to that of the phylogenetically inferred wobble position.

We assume all cassette exons represent a mixture of those under functional selection, and those produced by splicing errors (in the time scale of evolutionary processes, although they could be reproducibly detected in cells). We decompose the two groups by a Gaussian mixture model (GMM) using an expectation-maximization (EM) algorithm. More specifically, denote an exon *k* as $x_k = (s_i, s_w)$.

$$P(x_k \mid \Theta) = \alpha_n P(x_k \mid q_n) + \alpha_c P(x_k \mid q_c).$$

The first component represents noise, and the second component represents conserved AS events. $\theta = (\mu, \Sigma)$ denotes the mean and covariance matrix of exons in each group, and

$$P(x_k \mid \theta) = \frac{1}{2\pi |\Sigma|^{1/2}} e^{-\frac{1}{2}(x_k - \mu)^T \Sigma^{-1} (x_k - \mu)}$$

When we fit the data, we fixed θ_n ($\mu_n = [0.107, 0.698]$, and $\Sigma_n = \begin{bmatrix} 0.054 & 0.057 \\ 0.057 & 0.261 \end{bmatrix}$), which is estimated from constitutive exons. Parameters of the second component θ_c is estimated by EM ($\mu_c = [0.616, 0.767]$, and $\Sigma_c = \begin{bmatrix} 0.313 & 0.326 \\ 0.326 & 0.896 \end{bmatrix}$). The prior is estimated to be α_n=*0.704*, and α_c=*0.296*.

Each exon is then ranked by the posterior probability:

$$P(n \mid x_k, \theta) = \frac{\alpha_n P(x_k \mid \theta_n)}{\alpha_n P(x_k \mid \theta_n) + \alpha_c P(x_k \mid \theta_c)}$$

and the false discovery rate (FDR) of top *M* predictions is $\sum_{k=1}^{M} P(n \mid x_k, \theta) / M$ by definition.

Results

Mapping of the RNA-Seq reads

For the 2x50 reads, 78.3% can be mapped to the exons or exon junctions uniquely, while for 1x75 reads, the percentage is 80.2%. The reads from 16-tissue mixture (1x100 reads) have low mapping rate (30.0%), partly due to the short insert lengths, which are frequently shorter than the read size. Among the uniquely mapped reads, 9.33%, 17.4% and 31.5% are exon junction reads for 2x50, 1x75 and 1x100 reads, respectively (Supplementary Table 1). This agrees on the fact that longer reads have more chance to cross exon junctions. After merging the reads from all 36 experiments, totally 758,281 junctions were identified, including 488,356 novel junctions.

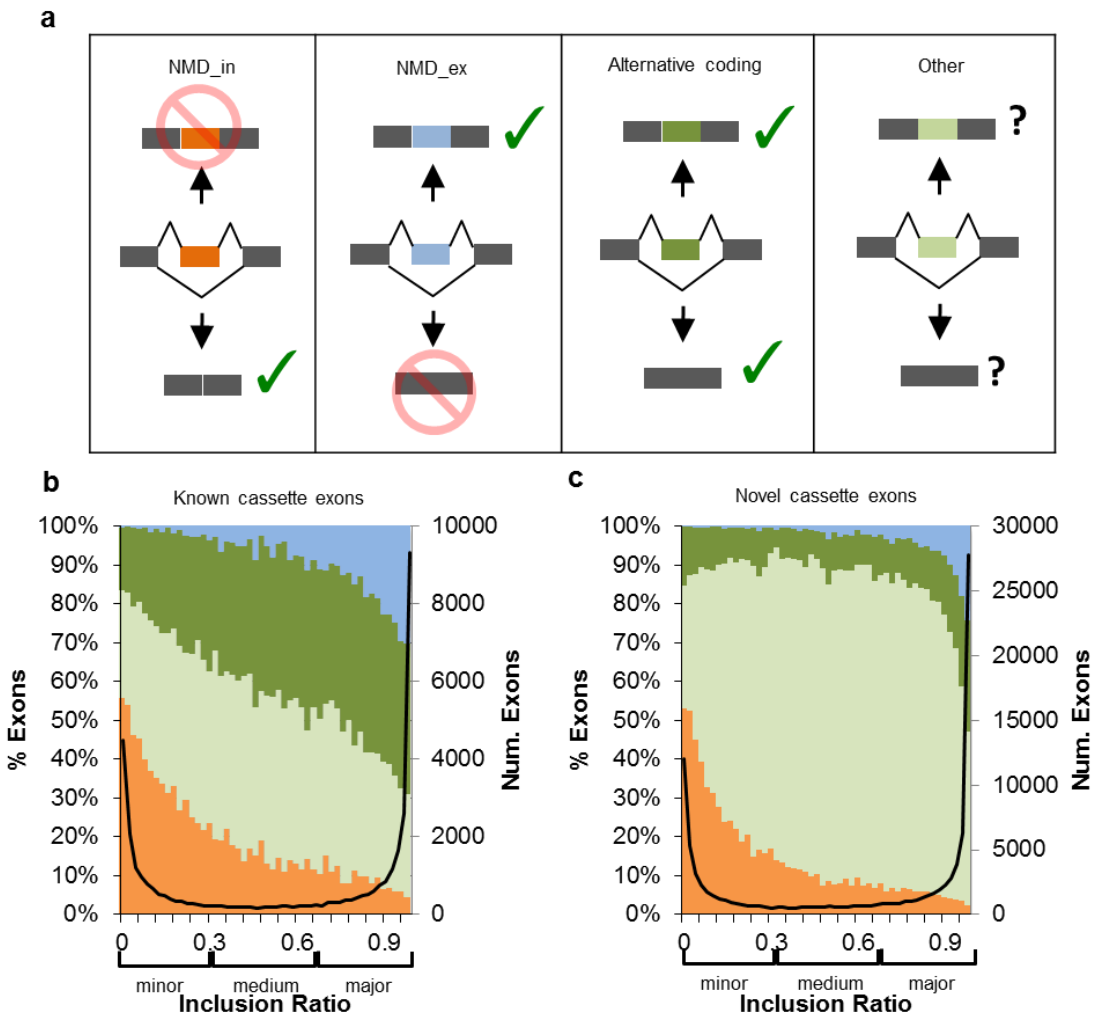


Figure 1. Classification of alternative splicing events based on protein coding (cassette exons). (a). Possible outcomes of alternative exons in terms of protein coding. NMD_in: transcript is degraded by NMD upon exon inclusion; NMD_ex: transcript is degraded by NMD upon exon exclusion; Alternative coding: both isoforms produce protein products; Other: All other exons that cannot be classified into the first three categories (ambiguous, or no ORF-containing transcripts observed). (b,c). Inclusion level of known and novel cassette exons observed in the data. Only those in annotated protein coding genes were included. The events which passed SpliceTrap cutoffs were used here and the inclusion ratios were averaged across different tissues. Exons were binned according to the inclusion level. For each bin, the percentage of exons in each category was plotted in different colors (left axis). The total number of exons in each bin was shown in black curve (right axis). The exons were then grouped into three catergories according to the inclusion ratios (0~0.33, 0.33~0.67, 0.67~1).

Identification of exons and AS events

After combining the junction reads and previous evidence of AS events in RefSeq transcripts, mRNAs and ESTs, we identified 120,110 cassette exons in the human genome, including 45,687 exons without previous evidence of inclusion and 36,841 exons without previous evidence of exclusion. Strikingly, we identified 3,069 exons smaller than 30 nt, of which 2,091 are novel. These include 105 exons with a length of 6 nt. Distribution of these cassette exons can be found in Supplementary Figure 1.

Discovery of NMD-triggering and alternative coding AS events

33.6% of the annotated cassette exon events are predicted to trigger NMD, which is consistent with previous studies (106), although our dataset include many folds of more events. For novel cassette exon events identified from the data, 23.1% were predicted as NMD triggering events.

We then examined a subset of known and novel cassette exons for which their inclusion ratios can be estimated reliably, according to the dynamic cutoff from SpliceTrap. We also required the exons to be located in protein coding genes. For both known and novel exons, we observed a lower abundance of the NMD isoforms (Figure 1).

Identification of AS events under strong selection pressure

A powerful method to evaluate the functional significance of AS events in a genome-wide scale is conservation. Conserved alternative splicing events are associated with a higher level of sequence conservation in the alternative exon and flanking intronic sequences, suggesting the selection pressure to preserve *cis* splicing elements (RBP binding sites, secondary RNA structures, etc) that are important for splicing regulation.

Computational methods have been used to predict exons with conserved splicing using sequence conservation levels, reading-frame preservation, exon and intron size, splice site strength, and additional sequence features (120-122). These methods typically used pairwise alignment (human and mouse) to measure sequence conservation, which is limited in statistical power, and also used features that favor protein-coding exons. Another study (123) detect the ratio of synonymous mutations versus non-synonymous

mutations in exons as a measure of selection pressure for splicing regulation, which is unsuitable for noncoding exons, and also has limited statistical power, because it did not consider signals in introns.

We note that different groups of cassette exons have different conservation features, when we examine a subset of exons which present in both human and mouse (Figure 2). Specifically, for NMD_in cassette exons, we also filter out those overlapping with protein coding exons, and thus have no selection pressure for protein-coding. The existence of orthologous sequences is very suggestive for their function. Indeed these exons are associated with the most elevated level of conservation in flanking introns, as measured by 46-way phastCons scores, when they are compared with exons spliced constitutively in human and mouse, and alternative coding exons in which both isoforms are relatively abundant, or those with relatively low inclusion level. In contrast, for NMD_ex exons, it is more difficult to distinguish them from constitutive exons due to the superimposition of selection pressure for protein coding and splicing. Nevertheless a higher conservation level is also observed when the exclusion isoform (targets of NMD) is relatively abundant. We also examined the conservation of the three codon position separately, using 46-way phyloP score (which measures the conservation of individual bases, as opposed to phastCons scores, which take neighboring sequences into consideration), and observed a lowest purifying selection pressure in the wobble positions of constitutive exons and NMD_ex_major, coding_major exons (see definition of major, medium, minor in Figure. 1), and higher conservation in the wobble position of NMD_ex_minor, NMD_ex_medium, coding_medium, and coding_minor, consistent with and extend previous studies (115,123,124).

The similarity of coding_major, NMD_ex_major, and constitutive exons suggest that as the sequencing depth increases in human and mouse, aberrant splicing products (the rare exclusion isoform) can be sampled and sequenced by chance in both human and mouse, so that the apparent conservation of splicing pattern is not a sufficient indicator of their functional importance (114).

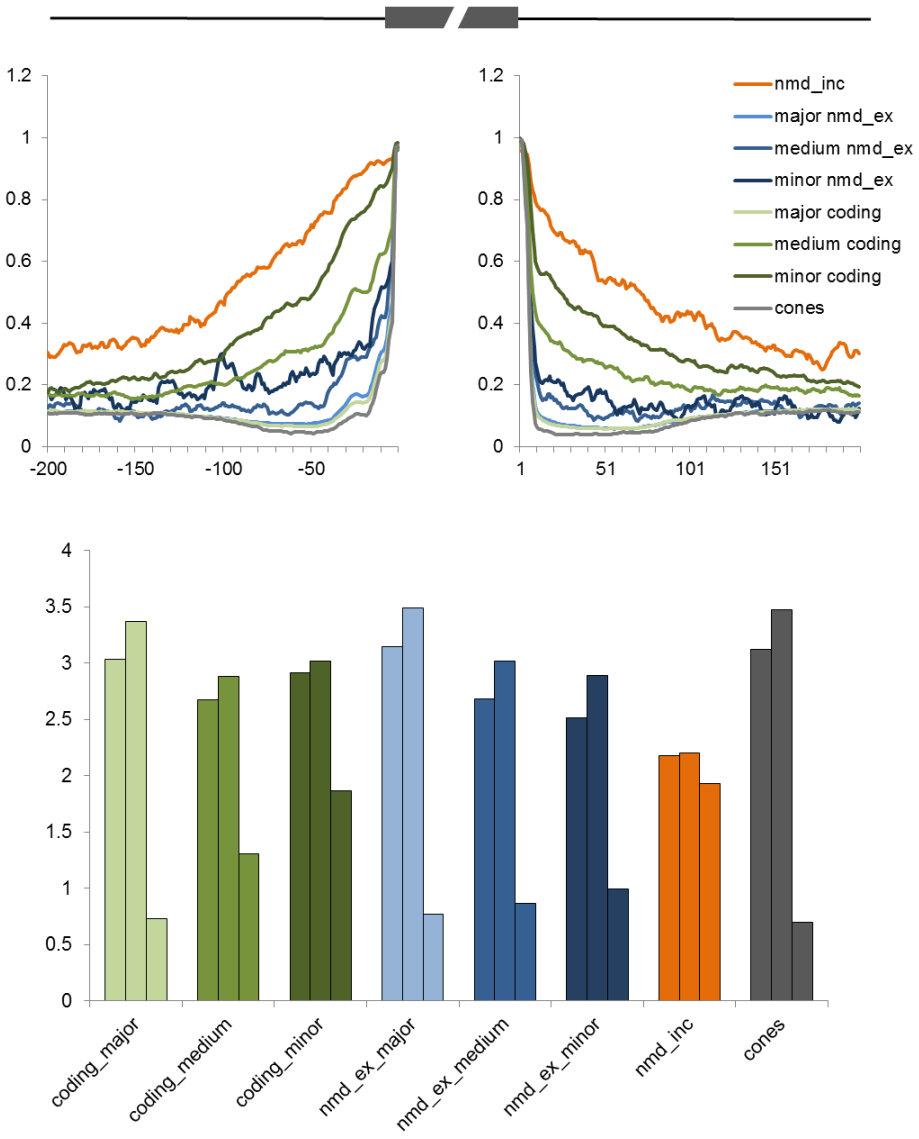


Figure 2. Sequence conservation of ancestral cassette exons. (a) Conservation profile (46 way phastCons scores) in 30 nt exonic sequences and 200 nt intronic sequences flanking the 3' and 5' splice sites of different groups of cassette exons. Alternative coding exons and NMD_ex exons are included only if the inclusion level can be estimated robustly from mRNA-Seq data and they are grouped further by their inclusion ratio (see Fig. 1b). NMD_in exons are required to have no overlap with any coding sequences from mouse, human and rat (coding sequences as defined by RefSeq and UCSC known genes), so that the included NMD_in exons have no section pressure for protein coding. Constitutive exons are those constitutively spliced in both mouse and human (≥20 supporting transcripts in each species, but no alternative splicing observed in this region). (b) The average conservation (46-way phyloP score) of the three different codon positions. For each exon the reading frame was phylogenetically inferred from the strength of selection, and the codon position with the smallest phyloP score is considered as the wobble position.

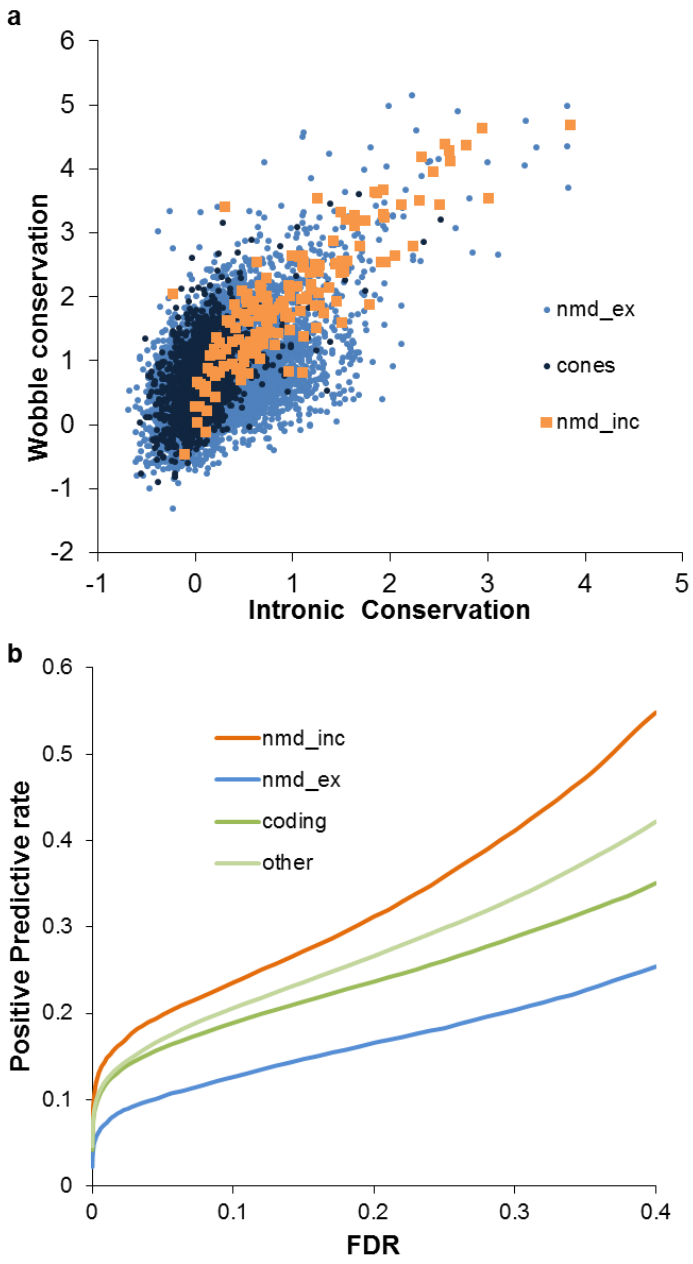


Figure 3. Sequence conservation to predict exons under significant selection using a Gaussian mixture model. (a) Scatter plot of ancestral NMD_ex, noncoding NMD_in, and constitutive exons (the same exons shown in Figure 2). X-axis is the average conservation score (46way phyloP score) of upstream and downstream introns. Y-axis is the conservation score of the phylogenetically inferred wobble position. (b) The number of selected exons defined by varying FDR reflecting varying stringency of conservation. Each category of exons was plotted separately. At FDR<0.05, we predicted 14,782 cassette exons under strong selection.

We find the conservation in the wobble positions in exons and flanking intronic sequences (200 nt each side) in many different species provide sufficient discriminative power to distinguish individual exons under strong selection to preserve regulated alternative splicing and those representing random sampling. Importantly, this method is suitable to measure selection pressure in exons whose sequences are subject to superimposed selection pressure for protein-coding and those that are not. We also avoided using features related to splicing, but do not reflect selection pressure (like exon size, splice site score, splicing regulatory elements).

Figure 3a shows scatterplots of noncoding NMD_in exons (positive control), which are largely separated from constitutive exons (negative control). NMD_ex exons are a mixture of both populations. We therefore studied all 91,010 cassette exons with conserved splice sites between human and mouse using a two-component Gaussian mixture model (GMM), assuming each exon belongs to either a population of exons under significant selection pressure to maintain regulated alternative splicing, or those that are not (similar to constitutive exons *per se*). Using this method, we estimated that 29.6% of all cassette exons are under selection. At the false discovery rate (FDR) of 5%, we were able to predict 14,782 cassette exons (16.2%) under significant selection pressure. This set included 94 of 161 (58.3%) stringently defined ancestral noncoding NMD_in exons. Overall, it included 2,554 NMD_in exons (1,663 are noncoding), 1,363 NMD_ex exons, 3,657 alternative coding exons, and 7,208 remaining exons for which we cannot reliable classify according to sequenced transcripts in the mouse genome.

Evolutionary history of NMD and alternative coding exons

We examined the origin of NMD_in exons (only those without overlapping coding exons), NMD_ex exons, and alternative coding exons, using constitutive exons as a control. The exons under strong selection are mostly conserved in mammals, due in part to the stringency of selection pressure we require (Figure 3). However, the pattern of conservation shows interesting difference between different groups. A majority of the noncoding NMD_in exons are conserved only in mammals, with some extending into non-mammal vertebrates. A small group of exons are conserved in all vertebrates. In contrast, for NMD_ex exons, a majority of exons are conserved in all vertebrates and a

small set of exons are missing in non-mammal vertebrates or fish. This suggested that the NMD_in exons were created during and before mammals and other vertebrate species split from the common ancestors (Figure 4)

Conclusions

This Chapter demonstrated an application of our pipeline on BodyMap 2.0 data, in which we discovered more than 80,000 novel cassette exon events. We further characterized the exons using a GMM model combining conservation features in both exons and flanking introns, and predicted a subset of conserved cassette exons under purifying selection pressure driven my splicing regulation.

This study also discovered more than 3,000 micro-exons smaller than 30 nt, including 105 exons as short as 6 nt. Compared to previous study of micro-exons using cDNA sequences, in which 170 novel micro-exons (≤25 nt) were detected in human (55), this study greatly extends our knowledge about this particular group of exons. Due to minimum information encoded in the exons, they can be used as excellent models for intronic splicing regulation study, and their functions also need careful examination.

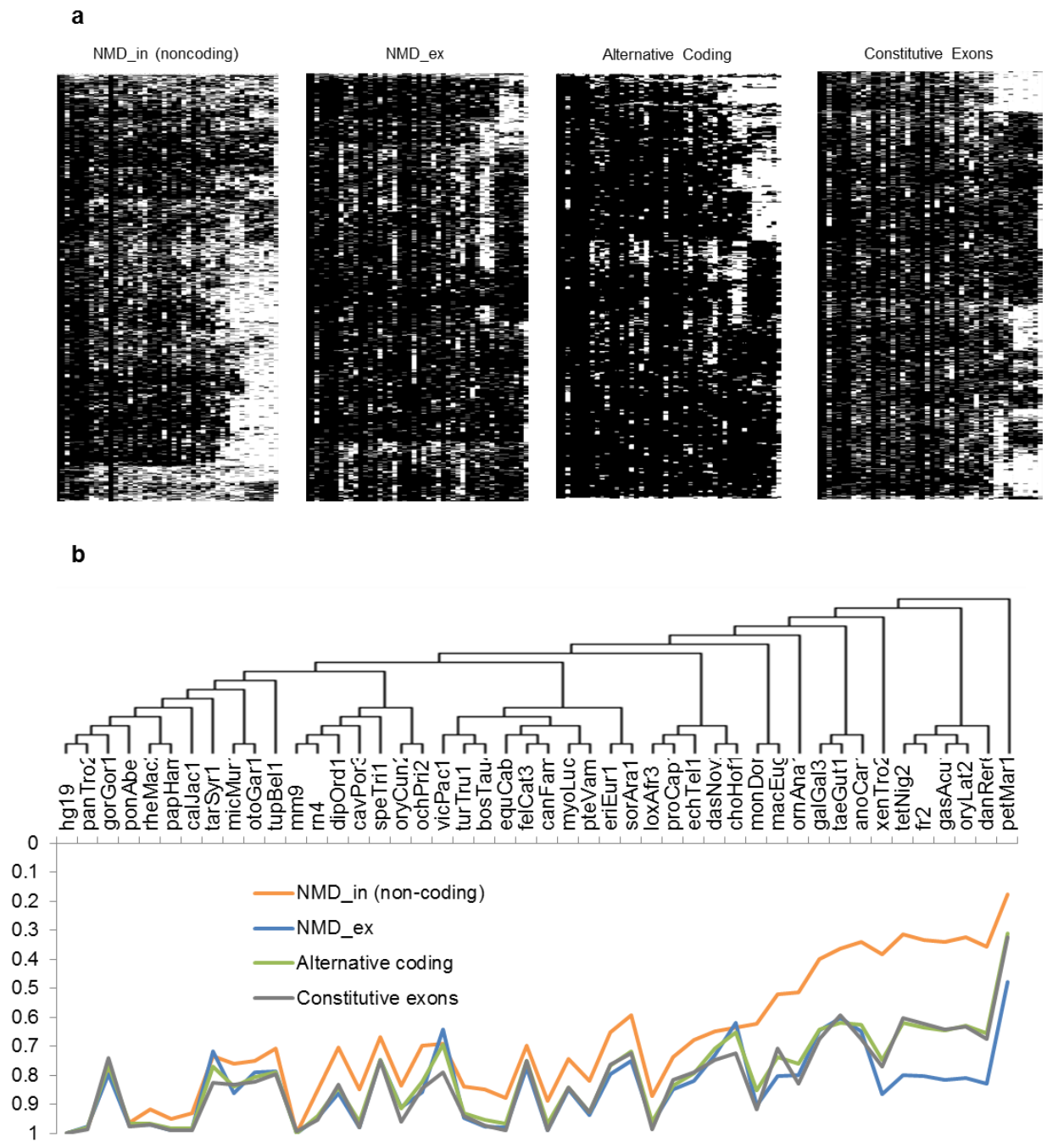
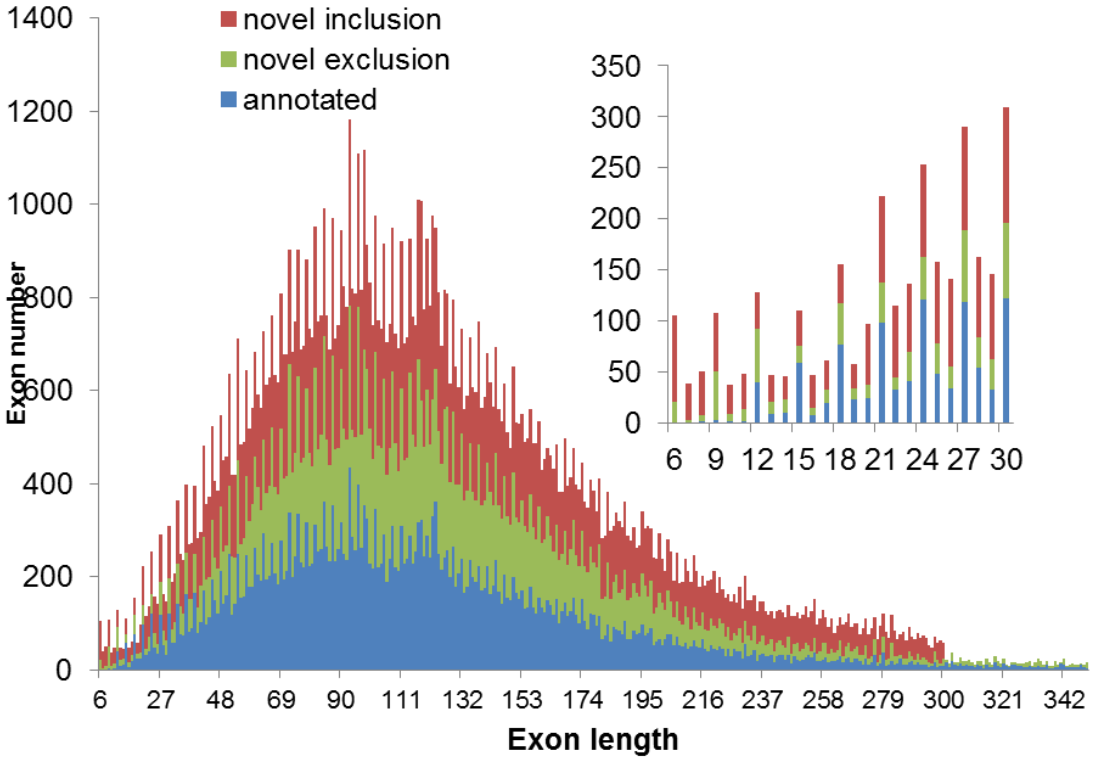


Figure 4. Evolutionary history of different groups of alternative and constitutive exons. (a) Heatmap showing conservation of each exon in each of the 46 vertebrate species (black means presence of the exon). (b) The percentage of conserved exons in each of the species, together with the phylogenetic tree of the species.

Supplementary Information



Supplementary Figure 1. Distribution of exon lengths. Blue color indicates that the exons are annotated cassette exons, green color indicates that the exons are annotated as constitutive exons, but in this study are defined as cassette exons, red color means the exons are novel cassette exons without any previous annotation.

Supplementary Table 1. Statistics of the BodyMap 2.0 data

ID	Tissue	Protocol	Stranded	Number of reads	Uniquely mapped reads	Uniquely mapped Junction reads
ERR030872	thyroid	Paired end	No	81912887	133256910	14223563
ERR030873	testes	Paired end	No	81836199	131902636	13854685
ERR030874	ovary	Paired end	No	80946260	129998453	12266841
ERR030875	whilte blood cells	Paired end	No	81217148	131638941	14515932
ERR030876	skeletal muscle	Paired end	No	82111139	126810111	13568084
ERR030877	prostate	Paired end	No	82334076	134915700	13118461
ERR030878	lymph node	Paired end	No	82078157	120943906	10760627
ERR030879	lung	Paired end	No	79296905	127582952	12013686
ERR030880	adipose	Paired end	No	77300072	118104527	10876840
ERR030881	adrenal	Paired end	No	74472871	116990320	9200490
ERR030882	brain	Paired end	No	73513047	116230657	8402795
ERR030883	breast	Paired end	No	75862215	117327162	9718513
ERR030884	colon	Paired end	No	82437443	121308021	10812500
ERR030885	kidney	Paired end	No	80397337	121617424	9297878
ERR030886	heart	Paired end	No	82918784	127337438	10658561
ERR030887	liver	Paired end	No	80048623	127044947	13531303
ERR030888	adipose	Single end	No	76269225	59522642	10168932
ERR030889	adrenal	Single end	No	76171569	61065046	9118598
ERR030890	brain	Single end	No	64313204	51975049	7106826
ERR030891	breast	Single end	No	77195260	61050977	9398524
ERR030892	colon	Single end	No	80257757	59759122	9937990
ERR030893	kidney	Single end	No	79772393	61151286	8532407
ERR030894	heart	Single end	No	76766862	58501236	8980570
ERR030895	liver	Single end	No	77453877	62260763	12286525
ERR030896	lung	Single end	No	81255438	67272296	11728192

ERR030897	lymph node	Single end	No	81916460	63737795	10743915
ERR030898	prostate	Single end	No	83319902	69938426	12786295
ERR030899	skeletal muscle	Single end	No	82864636	67087458	13061637
ERR030900	whilte blood cells	Single end	No	82785673	69060333	14375482
ERR030901	ovary	Single end	No	81003052	66580199	12063690
ERR030902	testes	Single end	No	82044319	67768811	13101912
ERR030903	thyroid	Single end	No	80246657	66850153	13145701
ERR030856	16 tissue mixture	Single end	Yes	76447153	28045330	9018894
ERR030857	16 tissue mixture	Single end	Yes	78243019	28279244	9162686
ERR030858	16 tissue mixture	Single end	Yes	77229855	27612881	8943117
ERR030859	16 tissue mixture	Single end	Yes	76274508	30287933	10178253
ERR030860	16 tissue mixture	Single end	Yes	75929029	30427363	10206931
ERR030861	16 tissue mixture	Single end	Yes	74756517	29860648	9964336
ERR030862	16 tissue mixture	Single end	Yes	73420952	11868079	3121051
ERR030863	16 tissue mixture	Single end	Yes	73520276	11812901	3108561
ERR030864	16 tissue mixture	Single end	Yes	77258890	27999364	9015398
ERR030865	16 tissue mixture	Single end	Yes	75982104	26952093	8638523
ERR030866	16 tissue mixture	Single end	Yes	74249497	28570554	9454276
ERR030867	16 tissue mixture	Single end	Yes	73773895	28399493	9374281
ERR030868	16 tissue mixture	Single end	Yes	72451624	11927186	3094006
ERR030869	16 tissue mixture	Single end	Yes	70743680	11812056	3057393
ERR030870	16 tissue mixture	Single end	Yes	71937539	11671839	3049014
ERR030871	16 tissue mixture	Single end	Yes	72321018	11748662	3063826

Chapter 6 :

Conclusions and Perspectives

This dissertation described the major work I have done in my Ph.D. study: a set of programs to detect, quantify and characterize alternative splicing from RNA-Seq data. This framework includes a sensitive mapping program (OLego) to align the reads to reference genome at a high resolution, a set of scripts to extract alternative splicing events from the mapping results, and an alternative splicing quantification method (SpliceTrap) to measure the splicing events in the data. All the programs are specifically designed for splicing study, hence the sensitivity and accuracy were particularly optimized for this specific biological problem. For example, OLego uses ultra-small seeds which enable detection of exons and exon junctions at a high resolution; SpliceTrap analyzes the splicing events and estimates splicing ratios by looking at the reads around the local regions instead of the whole transcripts, such that the complexity is reduced largely.

These programs and methods have been tested in simulations and later applied to real problems. The analysis demonstrated the sensitivity of our mapping program, especially for the ultra-small exons. In the BodyMap 2.0 data, we identified thousands of micro-exons (<30 nt). We also validated 15 out of 15 (100%) novel micro-exons discovered from mouse retina with RT-PCR, with a high correlation of the inclusion

ratios between RNA-Seq and RT-PCR. Due to the limited information encoded in these exons, they can serve as an excellent model for alternative splicing regulation study, which could be a very exciting follow-up project. The analysis also revealed our ability to detect and measure alternative splicing events from the data. In the BodyMap 2.0 data, we discovered more than 120,000 cassette exons in the human genome, including >80,000 exons without previous evidence of alternative splicing.

An immediate follow-up work would be more analysis with real data to explore specific biological problems, including characterizing functions of RNA binding proteins. Krainer Lab has generated many RNA-Seq data to investigate the functions of different splicing factors, e.g., Fox2 and SRSF1. SpliceTrap identified hundreds of alternative splicing events regulated by these factors, for example, 98 out of 124 targets of SRSF1 were validated by RT-PCR in Krainer Lab. The whole pipeline can be applied to investigate the data in more details.

There are many aspects which can be improved in these programs and algorithms. Some of them are on the programming level, e.g. optimization of the speed and calibration of the parameters. Particularly, we have optimized the speed of OLego by using longer seeds with overlaps in a later version (v1.1.2), such that a similar sensitivity can be achieved compared to using shorter seeds, while the speed is greatly improved. For example, using 15 nt seeds with 1 nt overlap speeds up the alignment process for three times, compared to using 14 nt non-overlapping seeds. Some other improvements need more profound changes to the models. For instance, SpliceTrap can be extended in the future to quantify more complex alternative splicing patterns, like mutually exclusive exons, which involves more than 2 isoforms. This will require a more general TXdb annotation format and an extended statistic model to describe more types of AS events. It is also practical to extend the alternative splicing discovery scripts to assemble the transcripts present in the data. The alternative splicing events along the transcripts can be used as features to construct a linear system, which can be solved to find out the major transcripts. This “local-to-global” approach is able to combine information along the gene locus and result in a different view of isoform de-convolution.

In Chapter 2 and 4, we compared OLego and SpliceTrap with other splice mappers and quantification tools to evaluate the performances. For these comparisons, we selected the most popular and recognized tools. However, there are other algorithms which are worth to mention. For example, STAR (125) (Spliced Transcripts Alignment to a Reference) is an ultrafast algorithm for RNA-Seq data alignment. It achieves a high speed by using uncompressed suffix arrays, which also results in huge memory consumption. MISO (48) (Mixture-of-isoforms) is an algorithm employing a similar exon-centric concept to estimate the exon expression levels from RNA-Seq data, albeit with different model and heuristic compared to SpliceTrap. To explore more differences between the methods, systematic tests in both simulation and real data should be carried out in the future.

Next generation sequencing technique is evolving rapidly. Longer sequences and higher through-put can largely increase the complexity of the problems. These brought more challenges to algorithms analyzing the huge data. In addition to necessary adjustments of the parameters, some occasions need to be considered carefully. One of them is the increase of read length and fragment size. When the read length is around 100 nt, most of the reads are shorter than the exon trios/duos used in the model of SpliceTrap. However, when this length increases, more reads could be missed, because SpliceTrap maps the reads to the exon trios/duos instead of the transcripts. Similar problem will occur to fragment size as well, because longer fragments means it is more likely that only one end of the read can be mapped to the exon trios/duos, and the information from the other half of the read is discarded and is not used in the estimation. These problems can be solved by extending the exon trios by including the flanking regions, e.g., the immediate neighbor exons.

More components can be added into this framework. For example, differential splicing detection can be done after quantification of splicing events in different conditions. This can be done with either simple *t*-tests or more profound models. With the inclusion ratio of each event, tissue specific splicing events can also be identified, which can contribute to functional analysis. The study of AS should not be limited to a single type of data. Together with the diverse data from different platforms and techniques, such

as microarrays, ESTs and CLIP-Seq, more sophisticated models can be built to take advantage of the sea of data.

Another potential direction is to construct a comprehensive splicing events database based on RNA-Seq data. In the BodyMap 2.0 data, we identified >80,000 novel cassette exon events, the most high confidence ones can be collected into a database for later alignment of other data. In this way, the cumulated alternative splicing events, together with the novel junctions, can be used as annotations for alignments and evaluation of other data.

Reference

1. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.

2. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.

3. Collins, F.S., Morgan, M. and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, **300**, 286-290.

4. Chow, L.T., Gelinas, R.E., Broker, T.R. and Roberts, R.J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12**, 1-8.

5. Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.

6. Jamison, S.F., Crow, A. and Garcia-Blanco, M.A. (1992) The spliceosome assembly pathway in mammalian extracts. *Mol Cell Biol*, **12**, 4279-4287.

7. Sammeth, M., Foissac, S. and Guigo, R. (2008) A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, **4**, e1000147.

8. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**, 1413-1415.

9. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.

10. Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R. and Zhang, M.Q. (2011) SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, **27**, 3010-3016.

11. Graveley, B.R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197-1211.

12. Krecic, A.M. and Swanson, M.S. (1999) hnRNP complexes: composition, structure, and function. *Curr Opin Cell Biol*, **11**, 363-371.

13. Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*, **13**, 1631-1637.

14. Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D. *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol*, **2**, e4.

15. Chang, Y.F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, **76**, 51-74.

16. Maquat, L.E. (2005) Nonsense-mediated mRNA decay in mammals. *J Cell Sci*, **118**, 1773-1776.

17. Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*, **100**, 189-192.

18. Blencowe, B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37-47.

19. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.

20. Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A. and Johnson, J.M. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*, **40**, 1416-1425.

21. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nature Genet.*, **37**, 844-852.

22. Clark, T., Schweitzer, A., Chen, T., Staples, M., Lu, G., Wang, H., Williams, A. and Blume, J. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.*, **8**, R64.

23. Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713-714.

24. Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z. and Zhang, M.Q. (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**, 2841-2842.

25. Lin, H., Zhang, Z., Zhang, M.Q., Ma, B. and Li, M. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics*, **24**, 2431-2437.

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

27. Ferragina, P. and Manzini, G. (2000), *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, pp. 390.

28. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

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

30. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966-1967.

31. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-628.

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

33. Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, **38**, e178.

34. Zhang, Y., Lameijer, E.W., t Hoen, P.A., Ning, Z., Slagboom, P.E. and Ye, K. (2012) PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics*, **28**, 479-486.

35. Au, K.F., Jiang, H., Lin, L., Xing, Y. and Wong, W.H. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*, **38**, 4570-4578.

36. Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.W., Peng, Z. and Yiu, S.M. (2011) SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Front Genet*, **2**, 46.

37. Dimon, M.T., Sorber, K. and DeRisi, J.L. (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One*, **5**, e13875.

38. Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873-881.

39. Wu, J., Anczukow, O., Krainer, A.R., Zhang, M.Q. and Zhang, C. (2013) OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res*, **41**, 5149-5163.

40. Garber, M., Grabherr, M.G., Guttman, M. and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*, **8**, 469-477.

41. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-515.

42. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, **28**, 503-510.

43. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods*, **7**, 909-912.

44. Li, W., Feng, J. and Jiang, T. (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J Comput Biol*, **18**, 1693-1707.

45. Li, J.J., Jiang, C.R., Brown, J.B., Huang, H. and Bickel, P.J. (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci U S A*, **108**, 19867-19872.

46. Xia, Z., Wen, J., Chang, C.C. and Zhou, X. (2011) NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics*, **12**, 162.

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

48. Katz, Y., Wang, E.T., Airoldi, E.M. and Burge, C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, **7**, 1009-1015.

49. Pachter, L. (2011) Models for transcript quantification from RNA-Seq.

50. Nilsen, T.W. and Graveley, B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457-463.

51. Cooper, T.A., Wan, L. and Dreyfuss, G. (2009) RNA and disease. *Cell*, **136**, 777-793.

52. Licatalosi, D.D. and Darnell, R.B. (2006) Splicing regulation in neurologic disease. *Neuron*, **52**, 93-101.

53. Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B. and Pierce, E.A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518-2528.

54. Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.-W., Peng, Z. and Yiu, S.-M. (2011) SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. *Frontiers in Genetics*, **2**.

55. Volfovsky, N., Haas, B.J. and Salzberg, S.L. (2003) Computational discovery of internal micro-exons. *Genome Res*, **13**, 1216-1221.

56. Ferragina, P. and Manzini, G. (2000) Opportunistic data structures with applications. *Proc FOCS 2000*, 390-398.

57. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.

58. Hastings, M.L. and Krainer, A.R. (2001) Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol*, **13**, 302-309.

59. Zhang, M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum Mol Genet*, **7**, 919-932.

60. Pertea, M., Lin, X. and Salzberg, S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*, **29**, 1185-1190.

61. Dogan, R.I., Getoor, L., Wilbur, W.J. and Mount, S.M. (2007) SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Res*, **35**, W285-291.

62. Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-p., Baldi, P.F. and Hertel, K.J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. USA*, **102**, 16176-16181.

63. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res*, **40**, D84-90.

64. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, **32**, D493-496.

65. Zhang, C., Hastings, M.L., Krainer, A.R. and Zhang, M.Q. (2007) Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *Proc. Natl. Acad. Sci. USA*, **104**, 15028-15033.

66. Burset, M., Seledtsov, I.A. and Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364-4375.

67. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

68. Kodama, Y., Shumway, M. and Leinonen, R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, **40**, D54-56.

69. Stamm, S., Zhang, M.Q., Marr, T.G. and Helfman, D.M. (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucl. Acids Res.*, **22**, 1515-1526.

70. Carlo, T., Sierra, R. and Berget, S.M. (2000) A 5' Splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol Cell Biol*, **20**, 3988-3995.

71. Xing, Y. and Lee, C. (2006) Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nature Rev. Genet.*, **7**, 499-509.

72. Maquat, L.E. (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.*, **5**, 89-99.

73. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.

74. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.

75. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, **5**, 621-628.

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

77. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967-974.

78. Slater, G. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

79. Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J. and Darnell, R.B. (2010) Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science*, **329**, 439-443.

80. Li, Y., Li-Byarlay, H., Burns, P., Borodovsky, M., Robinson, G.E. and Ma, J. (2013) TrueSight: a new algorithm for splice junction detection using RNA-seq. *Nucleic Acids Res*, **41**, e51.

81. Hiller, D., Jiang, H., Xu, W. and Wong, W.H. (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics*, **25**, 3056-3059.

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

83. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat Genet*, **30**, 29-30.

84. Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236-243.

85. Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A. and Johnson, J.M. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*, **40**, 1416-1425.

86. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141-2144.

87. Wang, E.T., Sandberg, R., Luo, S.J., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.

88. Relogio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R.B. and Valcarcel, J. (2005) Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*, **280**, 4779-4784.

89. Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, **37**, 844-852.

90. Baumer, D., Lee, S., Nicholson, G., Davies, J.L., Parkinson, N.J., Murray, L.M., Gillingwater, T.H., Ansorge, O., Davies, K.E. and Talbot, K. (2009) Alternative splicing events are a late feature of pathology in a mouse model of spinal muscular atrophy. *PLoS Genet*, **5**, e1000773.

91. Gupta, S., Zink, D., Korn, B., Vingron, M. and Haas, S.A. (2004) Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics*, **20**, 2579-2585.

92. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res*, **14**, 1617-1623.

93. Xie, H., Zhu, W.Y., Wasserman, A., Grebinskiy, V., Olson, A. and Mintz, L. (2002) Computational analysis of alternative splicing using EST tissue information. *Genomics*, **80**, 326-330.

94. Clark, T.A., Sugnet, C.W. and Ares, M., Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**, 907-910.

95. Fullwood, M.J., Wei, C.L., Liu, E.T. and Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*, **19**, 521-532.

96. Jiang, H. and Wong, W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032.

97. Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11**, R86.

98. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-65.

99. Zhang, C., Hastings, M.L., Krainer, A.R. and Zhang, M.Q. (2007) Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *Proc Natl Acad Sci U S A*, **104**, 15028-15033.

100. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-515.

101. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2009) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-515.

102. Peng, T., Xue, C., Bi, J., Li, T., Wang, X., Zhang, X. and Li, Y. (2008) Functional importance of different patterns of correlation between adjacent cassette exons in human and mouse. *BMC Genomics*, **9**, 191.

103. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, **268**, 78-94.

104. Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. (2010) Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet*, **6**, e1001236.

105. Kan, Z., States, D. and Gish, W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837-1845.

106. Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. USA*, **100**, 189-192.

107. de Lima Morais, D.A. and Harrison, P.M. (2010) Large-Scale Evidence for Conservation of NMD Candidature Across Mammals. *PLoS ONE*, **5**, e11695.

108. Hansen, K.D., Lareau, L.F., Blanchette, M., Green, R.E., Meng, Q., Rehwinkel, J., Gallusser, F.L., Izaurralde, E., Rio, D.C., Dudoit, S. *et al.* (2009) Genome-Wide Identification of Alternative Splice Forms Down-Regulated by Nonsense-Mediated mRNA Decay in <italic>Drosophila</italic>. *PLoS Genet*, **5**, e1000525.

109. Zhang, C., Krainer, A.R. and Zhang, M.Q. (2007) Evolutionary impact of limited splicing fidelity in mammalian genes. *Trends Genet*, **23**, 484-488.

110. Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291-336.

111. Baker, B.S. (1989) Sex in flies: the splice of life. *Nature*, **340**, 521-524.

112. Rehwinkel, J., Raes, J. and Izaurralde, E. (2006) Nonsense-mediated mRNA decay: target genes and functional diversification of effectors. *Trends in Biochemical Sciences*, **31**, 639-646.

113. Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J. and Blencowe, B.J. (2006) Quantitative microarray profiling provides evidence against

widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.*, **20**, 153-158.

114. Baek, D. and Green, P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl. Acad. Sci. USA*, **102**, 12813-12818.

115. Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631-1637.

116. Sugnet, C.W., Srinivasan, K., Clark, T.A., Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D. *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Computat. Biol.*, **2**, e4.

117. Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E. and Ares, M., Jr. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708-718.

118. Kozak, M. (1989) The scanning model for translation: an update. *The Journal of Cell Biology*, **108**, 229-241.

119. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110-121.

120. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci. USA*, **102**, 2850-2855.

121. Dror, G., Sorek, R. and Shamir, R. (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, **21**, 897-901.

122. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R. (2004) A Non-EST-Based Method for Exon-Skipping Prediction. *Genome Res.*, **14**, 1617-1623.

123. Lu, H., Lin, L., Sato, S., Xing, Y. and Lee, C.J. (2009) Predicting Functional Alternative Splicing by Measuring RNA Selection Pressure from Multigenome Alignments. *PLoS Comput Biol*, **5**, e1000608.

124. Xing, Y. and Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA*, **102**, 13526-13531.

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