A

Splice Junction Discovery in RNA-Seq Data

AHMAD TAVAKOLI
University of Windsor
tavakola@uwindsor.ca

RNA-Seq has become the method of choice to study gene expression and has been dubbed as a revolutionary tool to study transcriptomics, and it has been used to detect splice junctions since its first days. Splice Junctions are the key to conduct research on alternative splicing, which is related to some complex human diseases. In this article, we review research on methods for finding splice junctions in RNA-Seq datasets. We come to the conclusion that there is a high similarity between the approaches reviewed in this survey. Along with improving the sensitivity and specificity of these approaches, computing power and memory consumption optimization is the other way of progress for the splice junction discovery approaches.

General Terms: Splice Junction, RNA-Seq, Filtering, Read counting, next-generation sequencing, spliced mapping, exon-intron boundary, support vector machines, hidden Markov model, maximum likelihood estimation.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Contents

1. INTRODUCTION

Detecting splice junctions has always been one of the interesting fields of studying genomics, and microarrays have been used extensively for this purpose in the past. Since 2008, next-generation sequencing has become the prominent method to study transcriptomics. This survey focuses on the methods for finding splice junctions in RNA-Seq data. The research papers for this survey were found using Google Scholar. We have found 21 papers, which studied the problem of splicing in RNA-Seq on different levels. 10 papers which tackled the problem of splice junction discovery directly and were published in most-cited journals were chosen.

The papers reviewed in this survey have been categorized according to the method that the authors have developed to detect splice sites. In the first category, we present the method developed by De Bona et al. [2008], Dimon et al. [2010], and Lou et al. [2011a]. The authors in this section have used some sort of machine learning algorithm to in their approach. In the second section of the survey, we would study the methods which try to assess the reliability of a possible splice junction by a read-count method. This section consists of works by Trapnell et al. [2009], Wang et al. [2010], Huang et al. [2011], and Zhang et al. [2012]. The third section focuses on papers that used other methods as their main way of removing false positives from their results. These papers include those of Au et al. [2010], Ameur et al. [2010], and Bryant et al. [2010]. We conclude our observations of scholarly works completed on the subject of splice junction discovery in the last part of the survey.

RNA-Seq started to be used in research after 2008. As a result, all papers of this survey have been published over a time spanning less than four years. Researchers may not have enough time to study the works of their peers and make comparison and experiment on already developed methods. The works by Ameur et al. [2010], Au et al. [2010], and Bryant et al. [2010] which is studied in this survey, have been published within a month. Obviously, none of them had the chance to refer to each other. The interesting point is that these papers fell into the same section in this survey, which implies the similarity of their work.

2. SURVEY OF RESEARCH

2.1. Machine Learning Methods

The papers presented in this section are distinguished from others by incorporating a machine learning technique. De Bona et al. [2008] is the first paper to discuss the problem of finding splice junctions in RNA-Seq data which used support vector machines as its approach. Dimon et al. [2010] use a hidden Markov model in addressing this problem.Lou et al. [2011a] used maximum likelihood estimation to find the location of splice junctions.

2.1.1. Support Vector Machine Learning. Short reads acquired from high-throughput sequencing technologies can be used for studying transcriptome and gene structure identification. Aligning these reads over intron/exon boundaries is a requirement for this purpose. The authors do not specify the reason for studying transcriptome explicitly.

De Bona et al. [2008] do not refer to any previous work which relates to the subject of this survey.

De Bona et al. [2008] developed their approach, called QPALMA, in three independent parts, splice site prediction model, dynamic programming algorithm and scoring function. QPALMA aims to align short reads to the reference genome; splice site prediction is helping this approach to achieve better results. This part is based on a machine learning

approach which uses a set of donor and acceptor sites to train a SVM predictor. The authors propose three different extensions for Smith-Waterman algorithm for aligning the reads to the reference.

The authors state that they trained their algorithm using a simulated dataset of 10,000 previously aligned sequences. The alignment error rate for these rates has been calculated incorporating different available informations. The authors also tried their approach on a dataset of spliced and unspliced 2.98 million reads of forward strands of chromosome 1.

De Bona et al. [2008] claim that they could align 10,000 *in silico* spliced reads with an error rate of 1.78% incorporating quality information, intron length model and splice site predictions. The authors state that it was the best rate that they have achieved. The authors also claim that QPALMA aligned spliced and unspliced reads with a 5.2% and 1.2% error rate respectively.

The authors claim that they could successfully exploit all information sources to align short reads over exon boundaries. The authors claim that their approach works reasonably for all next-generation sequencing platforms, including *Illumina* sequencing, which have been tried in their experiment.

The authors state that their method can be extended to exploit homo-polymer errors, which is available for Roche's 454 sequencing platform.

2.1.2. Method based on Hidden Markov Model. During the past decade, there has been a growing appreciation of the importance of alternative splicing as a mechanism for organisms to increase proteomic diversity and regulatory complexity. According to authors, the ability to detect alternative splice isoforms with accuracy and sensitivity is the key to comprehensive RNA-Seq analysis.

The authors refer to previous work by Mortazavi et al. [2008], Trapnell et al. [2009], Bryant et al. [2010], and Ameur et al. [2010].

The authors note that the method developed by Mortazavi et al. [2008] does not address the question of novel junctions and cannot be used for organisms with incomplete or inaccurate genome annotations. They state that the algorithm developed by Trapnell et al. [2009] performs best on mammalian transcripts with relatively high abundance, but has defects in more compact genomes and with non-canonical junctions. They note that the method proposed by Ameur et al. [2010] has the requirement for at least one read to split evenly across the exon-exon boundary which reduces sensitivity in low coverage datasets and transcripts, also they claim that this method supports only ABI SOLiD reads.

Dimon et al. [2010] state that SuperSplat, the method developed by Bryant et al. [2010], requires both pieces of a read to be exact matches to the reference sequence and conclude that it is not robust against sequencing errors or single-nucleotide polymorphisms. The authors claim that the algorithm designed by Au et al. [2010] considers only canonical splice junctions and requires read lengths of 50bp or greater.

Dimon et al. [2010] claim that they have developed a method to avoid the inherent bias introduced by relying upon previously defined biological information. Their algorithm, called HMMSPlicer, works by dividing each read in half and seeding the read-halves against the genome and using a Hidden Markov Model to determine the exon boundary. They claim that both canonical and non-canonical junctions are reported and a score is assigned

to each junction, which is dependent on the strength of the alignment and the number and quality of bases supporting the splice junction.

The authors claim that they compared their algorithm with algorithms designed by Trapnell et al. [2009] and Au et al. [2010]. They state that they analyzed the performance of their method on simulated reads and three publicly available experimental datasets.

Dimon et al. include the detailed results of their experiments with different algorithm parameters on examined datasets. The authors state that in comparison with TopHat, HMMSplicer shows its ability to find more junctions with a similar level of specificity in each of tested datasets. They state that in comparison with SpliceMap by Au et al. [2010], their method achieves 7% more matching junctions for human datasets, and it outperforms SpliceMap in the low sequence quality *A. thaliana* dataset.

The authors claim that HMMSplicer combines high sensitivity with a low false positive rate, functions properly on datasets with low quality sequence reads, performs well in datasets with uneven coverage, identifies many junctions in low abundance transcripts and also identifies non-canonical junctions, and finds true novel junctions in genomes with incomplete annotation.

The authors claim that their algorithm is the only software package that provides a score for each junction, reflecting the strength of the junction prediction.

2.1.3. Method based on Maximum Likelihood Estimation. Studying the way that alternative splicing affects a biological system is as important as studying its fundamental regulatory mechanisms, and as RNA-Seq provides the ability to analyze transcriptome in a base-level resolution and high coverage.

Lou et al. [2011a] refer to the work by Mortazavi et al. [2008], Trapnell et al. [2009], Ameur et al. [2010], Au et al. [2010], Bryant et al. [2010], and Lou et al. [2011b].

The authors state that the approach presented by Trapnell et al. [2009] depends on the canonical splice site motifs. The authors also state that as the methods developed by Ameur et al. [2010], Au et al. [2010] and Bryant et al. [2010] designed based on the idea of read counting, sequencing depth can significantly affect their performance.

Lou et al. [2011a] proposed an approach based on maximum likelihood estimation, which relies on geometric-tail distribution of intron lengths for aligning of paired-end RNA-Seq reads. The authors used a package named ABMapper, which was particularly developed for spliced mapping by the same team as the authors and is explained in [Lou et al. 2011b]. They state that their approach is an empirical probabilistic model which adopted a two-part distribution, an arbitrary length distribution and a geometric distribution. This method uses maximum likelihood to estimate the most probable location for a paired-end read based on this two-part distribution. The authors stated that their approach works in three models, one without any *a priori* knowledge, and two with expression level and junction-site frequency as *a priori* knowledge.

The authors state that they compared their model with methods developed by Trapnell et al. [2009] and Au et al. [2010]. They used two human lymphoblastoid cell-line datasets for testing purpose. The dataset consist of 8.4 million 75 bp paired-end reads with an approximately 250 bp insert size. The results were validated with Alternative Splicing and

Transcript Diversity (ASTD) database and Human EST database.

Lou et al. [2011a] claim that their method could report 53% and 49% more splice junctions comparing to the methods by Au et al. [2010] and Trapnell et al. [2009]. The authors also claim that 60% of the junctions which were predicted only by their method could be validated by ASTD, which comprised 22% of the total reported splice junctions. According to the authors, this implies that the methods proposed by Au et al. [2010] and Trapnell et al. [2009] missed at least one-fifth of the true splice junctions. The authors claim that by performing an exhaustive search for junctions in Human EST database, they found that their method predicted splice junctions with an accuracy of 96%.

The authors claim that their proposed approach can detect 50% more splice junctions than other existing tools. Lou et al. [2011a] claimed that the reason for superiority of their approach is in using first, ABMapper, which has a much higher sensitivity in spliced-mapping than other approaches, and the second is the geometric-tail based model.

Year	Author	Package Name	Notes
2008	De Bona et al.	QPALMA	Cited by Trapnell et al. [2009], Bryant et al. [2010],Wang et al. [2010], and Huang et al. [2011] Last updated in December 2010. The package is freely available.
2010	Dimon et al.	HMMSplicer	Cited by Huang et al. [2011], and Zhang et al. [2012] Last updated in November 2010. The package is freely available.
2011	Lou et al.	N/A	Not Cited. The package is not available.

2.2. Counting-Based Filtering Methods

The papers reviewed in this section, use filtering methods based on counting number of reads covering reference genome. Papers presented by Au et al. [2010] and Ameur et al. [2010] maps the reads by splitting them. The method developed by Bryant et al. [2010] use also empirical data for supporting possible junctions.

2.2.1. Mapping reads by splitting. The works presented in this section describe the methods developed by Au et al. [2010] and Ameur et al. [2010]. Both of them first split the reads, then try to map the fragments into the reference genome.

High-throughput sequencing of mRNA opens extraordinary opportunities to identify the spectrum of splice events in a sample on a global scale. An important application of deep RNA sequencing is the discovery of fusion transcripts in cancer. Definite fusion transcripts are commonly produced by cancer cells, and detection of fusion transcripts is part of routine diagnostics of certain cancer types. Abnormal RNA splicing is associated with many human diseases. For this reason, methods to identify and quantify splicing events are important to biology and medicine.

Ameur et al. [2010] refer to the work by Trapnell et. al [2009].

The authors state that in the method designed by Trapnell et al. [2009], a substantial number of true splice junctions, including junctions with long introns or non-canonical splice sites are outside of the detection range, and also this method is computationally challenging for transcripts expressed at lower levels.

The authors state that their method consists of a combination of a split-read alignment and the novel SplitSeek program. The alignment is performed using the AB/SOLiD whole-transcriptome-alignment software. The proposed method by Ameur et al. [2010], SplitSeek, developed in a way to find junction reads in which as few as five bases overlap with the other exon. It finds exon-exon boundaries that are supported by several split reads, it is required for each junction to be covered by at least two reads with unique starting points.

The authors state that they evaluated their method using public RNA-seq data from single mouse oocytes, which was performed on two independent samples, which consist of 50-bp reads.

The authors state that they selected 22 base pair for the anchor length according to the highest number of uniquely mapped split reads was obtained for this length. Ameur et al. [2010] present the results of their experiment in terms of the number of splice junctions and insertions, number of predicted small insertions and deletions within RefSeq exons, and number of predicted splice junctions as a function of the total number of processed reads.

The authors claim that the exon-exon boundaries are identified almost at nucleotide resolution and with a low false-positive rate, less than one in 10,000, for junctions within 100 kb. Ameur et al. [2010] state that their method makes it possible to study splice junctions and fusion genes while also measuring the gene expression using RNA-Seq data. They claim according to their results, their proposed algorithm has a very low false-positive rate, and they state that acquired false discovery rate of less than one for 1,000 junctions within 1Mb and less than 1/10,000 for those within 100 kb, is supporting their claim.

Au et al. [2010], in their research paper, state that the method developed by Mortazavi et al. [2008] is dependent on annotated exon library and since the exon, library is incomplete, this method cannot find junctions that involve novel splicing events. The authors do not mention any shortcomings of the method presented by Trapnell et al. [2009].

Au et al. [2009] present their method, SpliceMap, based on the idea using the mapping of half-reads as a way to identify the approximate location of a junction. SpliceMap works in four steps, half-read mapping, seeding selection, junction search and paired-end filtering. It maps both halves of the read to the reference genome by a short read mapping tool.

The authors state that they compared their method with the method described by Trapnell et al. [2009] on an RNA-Seq dataset of 23,41,226 reads. They claimed that they assessed their method's specificity by aligning detected junctions to human ESTs in GenBank. They also stated that they investigated novelty of discovered junctions by PCR experiments. Au et al. [2010] describe a comparison with ERANGE method by Mortazavi et al. [2008]. They also stated that they compared their method with BLAT, which is a common tool for EST sequences alignment. The authors claimed that they calculated the performance of their method in a specific CPU running time and compared it with TopHat

by Trapnell et al. [2009].

Au et al. [2010] claim that 87.9% of junctions found by their method were supported by EST evidence. They state that SpliceMap achieves more than 95% sensitivity for highly expressed genes, more than 90% for genes with medium expression and (40-67%) for genes with low expression. They state that more genes detected by SpliceMap are of higher degree (80100%) of completeness in junction discovery. Au et al. [2010] claim that in a random sample experiment, 85% of novel junctions validated using PCR experiment.

The authors stated that ERANGE package by Mortazavi et al. [2009] found 160,899 junctions and SpliceMap found 151,317 junctions, among those found by [Au et al. 2010] method, 23,020 junctions, which were not found by ERANGE, were novel. They also claimed that the BLAT package, achieved a similar but still slightly lower level of specificity with a much lower sensitivity (70% lower) as compared to SpliceMap.

The authors stated that it took 66 CPU hours for SpliceMap and 12 CPU hours for ERANGE to process the data set.

They claim that based on their results, SpliceMap detects more annotated junctions than TopHat, method presented by Trapnell et al. [2009]. They claim that 50bp reads can support an approach of direct de novo detection of splice junctions without the need to first cluster reads to identify Bryant et al. [2010] introduce a new approach, called Supersplat, that uses hash table as a way to save system memory by storing read sequences as key and their frequencies as the value. Supersplat uses two parameters to limit the maximum and minimum length of the sequence which it is trying to build location indexes based upon them. After indexing the reference sequence, Supersplat identifies reads that can be aligned against the reference, as possible splice junctions in an iterative process. Potential splice junctions are filtered according to the number of overlapping reads on two intron boundaries.accepted exons, and that this approach can achieve significantly higher sensitivity in junction detection than current leading methods of RNA-Seq analysis. They also claim that paired-read information can help to reduce false discoveries.

2.2.2. Splice junction discovery using empirical RNA-Seq data. Next-generation sequencing provides a ground for massive transcript expression analysis. RNA-Seq approach supplies enough reads for this purpose, and the key to profiling this genetic information is identification of intron/exon boundaries or splice junctions.

Bryant et al. [2010] refer to the work by De Bona et al. [2008], Trapnell et al. [2009], and Filichkin et al. [2010] in their paper.

Bryant et al. [2010] state that the method developed by De Bona et al. [2008] relies on the previously known splice sites for training the algorithm which influences the results. Furthermore, they note that QPALMA scores junctions that conform with canonical splicing motifs higher, so it may be inefficient in finding non-canonical splice junctions. The authors state that this problem also applies to TopHat, the method developed by Trapnell et al. [2009]. Bryant et al. [2010] state that TopHat needs a high number of RNA-Seq reads to build exon islands.

Bryant et al. [2010] introduce a new approach, called Supersplat, that uses hash table as a way to save system memory by storing read sequences as key and their frequencies as the value. Supersplat uses two parameters to limit the maximum and minimum length of the sequence which it is trying to build location indexes based upon them. After indexing the reference sequence, Supersplat identifies reads that can be aligned

against the reference, as possible splice junctions in an iterative process. Potential splice junctions are filtered according to the number of overlapping reads on two intron boundaries.

The authors state that they tested the performance of their method on a set of 3,690,882 *Arabidopsis thaliana* reads. They used TAIR8 database of annotated junctions to evaluate Supersplat's performance. Bryant et al. [2010] also state that they assessed their approach for *de novo* splice junction discovery on a dataset of *Brachypodium distachyon*.

The authors claim that they confirmed 91% of canonical and 86% non-canonical splice junctions using PCR and Sanger sequencing. Bryant et al. [2010] claimed that they achieved a predicted positive rate (PPV) of 70% with the minimum read length of 6 and a 90% rate by increasing it. According to Bryant et al. [2010], this rate reach 97% by setting the overlapping number of reads filter to 21.

Bryant et al. [2010] claim that their approach is unbiased and exhaustive, but it may generate output files with up to tens of gigabytes size, and the user should account for determination of the befitting criterion to filter out spurious output. The authors claim that the exhaustive approach of their method can discover many previously unknown splice junctions.

Year	Author	Package Name	Notes
2010	Au et al.	SpliceMap	Cited by Wang et al. [2010], Dimon et al. [2010],Lou et al. [2011a], and Huang et al. [2011] Last update in October 2010. Source code is freely available.
2010	Ameur et al.	SplitSeek	Cited by Dimon et al. [2010], and Lou et al. [2011a] Not available for download. Not being maintained.
2010	Bryant et al.	Supersplat	Cited by Dimon et al. [2010], Lou et al. [2011a],Huang et al. [2011] Not available for download. Not being maintained.

2.3. Non-counting Filtering Methods

The common element between the papers presented in this subsection is using various filtering techniques to omit spurious splice junctions. As all of these methods use simple searching methods to find junctions, they need some sort of filtering to detect false positives and gain higher sensitivity. Although, they developed methods that used various strategies in filtering and also were highly similar to each other in nature. In this survey, We tried to be as much specific as possible in categorizing the papers presented in this section. The work developed by Trapnell et al. [2009] has been placed on filtering based on average read depth coverage section, the work of Wang et al. [2010] on filtering on minimum anchor length, and the works of Huang et al. [2011] and Zhang et al. [2012] on filtering using paired-end information and read coverage.

2.3.1. Filtering based on average read depth coverage. Alternative splicing is a significant process in normal cellular functions and also in human diseases. Finding novel splice junctions is an important part of studying alternative splicing.

Trapnell et al. [2009] refer to previous work by De Bona et al. [2008] and Mortazavi et al. [2008].

The authors mention two shortcomings of the work of De Bona et al. [2008], the first is that their method, QPALMA, depends on a set of known splice junctions from the reference genome and cannot identify novel junctions. They state that the other shortcoming is that De Bona et al. use Vmatch, an alignment program which is not designed to map short reads on machines with small main memories and is considerably slower than other short-read mappers. Trapnell et al. [2009] state that ERANGE, the method developed by Mortazavi et. al, depends on available annotation of exon-exon junctions for its main objective, which is gene expression quantification in mammalian RNA-Seq projects.

The authors introduce a new system called TopHat, which works in two phases to find junctions. In the first phase, all reads are mapped to the reference genome using Bowtie, all reads that do not map to the reference genome are set aside as initially unmapped reads. Then an initial consensus of mapped regions, called exon islands, is computed using the assembly module in a package named Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form prospective splice junctions. For each splice junction, Tophat searches the initially unmapped reads in order to find reads that span junctions using a seed-and-extend strategy.

Trapnell et al. [2009] state that they conducted an experiment on 47,781,892 short reads using their method, TopHat, and a previously developed method called ERANGE by Mortazavi et al. [2008].

The authors claim that their method could discover around 72% of splice junctions comparing to annotation-based analysis done in [Mortazavi et al. 2008] in fewer transcribed regions and 80% of junctions in more actively transcribed regions. They claimed that out of 19,722 newly discovered junctions that they found in their experiment, many of them are true splices, but it is difficult to asses exactly how many of them are genuine.

Trapnell et al. [2009] claim that the significance of their work is in its ability to detect novel splice junctions. They also claimed that their tool represents a significant advance over previous RNA-Seq splice detection methods.

2.3.2. Filtering on minimum anchor length. Accurate identification and quantification of transcript isoforms is crucial to characterize alternative splicing among different cell types. In addition, sequence variants found within splice sites or splicing enhancer sequences may have functional consequences on alternative splicing. A large proportion of human genetic disorders results from splicing variants.

Wang et al. [2010] refer to the work by De Bona et al. [2008], Trapnell et al. [2009] and Au et al. [2010]

The authors note that the output generated by the method of Au et al. [2010] does not include tag alignments, hence is incomplete. They do not state any shortcoming regarding the works of others.

According to Wang et al. [2010], their method operates in two phases. In the first phase, called tag alignment, candidate alignments of the mRNA tags to the reference genome are determined. A set of candidate alignments are computed for each tag as multiple possible alignments may be found for each read. In the second phase which is called splice inference phase, splice junctions that appear in the alignments of one or more tags are analyzed to determine a splice significance score based on the quality and diversity of alignments that include the splice. The most likely alignment for each tag is chosen based on the splice significance score.

The authors state that they evaluated specificity and sensitivity of their method using an experiment on a generated synthetic dataset. They also state that they validated their method using qRt-PCR experiment.

[Wang et al. 2010] state that they achieved a true-positive rate of 96% and false-positive rate of 8% for their method. They stated that over 77% of canonical junctions found by their method were confirmed by known transcripts in GenBank, which was between 6% to 11% higher in comparison by TopHat method by Trapnell et al. [2009].

The authors claim that both TopHat by Trapnell et al. [2009] and their method were more memory efficient and much faster in experiments than SpliceMap by Au et al. [2010]. They also claim that their method performed best by detecting more true-positive junctions and fewer false-positive junctions than the other two methods. They state that longer tags improve both the sensitivity and the specificity of the junction discovery in their method and as well in method by Trapnell et al. [2009], and they claim that in comparison, their method has a higher sensitivity in different tag lengths. They claim that using read lengths of 75 or 100bp yield to significantly better sensitivity and specificity for splice detection.

2.3.3. Filtering using paired-end information and read coverage. *Problem.* Splice junction detection is the first step of studying alternative splicing. Alternative splicing is highly effective on diversity of proteins, as it causes different mRNAs to be produced from the same gene. These different mRNAs translate into different protein isoforms.

Huang et al. [2011] refer to the previous work by Mortazavi et al. [2008], De Bona et al. [2008], Trapnell et al. [2009], Bryant et al. [2010], Au et al. [2010], Wang et. al [2010], and Dimon et al. [2010].

The authors state that QPALMA, the method developed by De Bona et al. [2008] which uses a machine learning approach, is biased toward splice junctions that are similar to the ones in the training data set. Huang et al. [2011] state that low sequencing depth affects the performance of the algorithm developed by Trapnell et al. [2009] and hence there would not be enough reads for efficient junction detection. The authors state that the method introduced by Bryant et al. [2010], which uses hashing as its alignment approach, needs a large amount of memory and computing power and as a result is not scalable for reads longer than 50 base pairs.

Huang et al. [2011] state that SpliceMap, the algorithm presented by Au et al. [2010], performs poorly while dealing with the reads that can be mapped to more than one location. Furthermore, they state that this approach is not efficient when the transcriptome is lowly expressed or the reads have sequencing errors. The authors state that the method developed by Wang et al. [2010] has some inefficiencies while the sequencing depth is low, which leads to reduced call rate.

Huang et al. [2011] present SOAPsplice, which finds the splice junctions in two steps. In the first step, it maps the reads into reference genome using Burrows Wheeler Transformation for indexing. Then SOAPSplice detects splice junction candidates based upon some criteria, which include following known splicing motifs and a maximum intron size of 50,000 bp. SOAPsplice applies two different filtering techniques to omit false positives. The first strategy is to check the paired-end information with the direction of the mate-pair reads and later discarding incompatible junctions. The other strategy is to filter out the junctions that have a missing segment between two sub-reads that have been mapped to the reference genome.

Huang et al. [2011] compare their method with the algorithms developed by Trapnell et al. [2009], Wang et al. [2010], and Au et al. [2010] on two 50 and 150 bp simulated dataset and two 51 and 130 bp real dataset.

The authors claim that according to the results of the simulated dataset for both 50 and 150 bp length reads, their method had the highest call rate while it kept the false positive rate at its lowest comparing to other approaches. For the real dataset with 51 bp reads, Huang et al. [2011] claim that SOAPsplice detects more novel junctions than TopHat by Trapnell et al. [2009] and its results are comparable to the method designed by Au et al. [2010] on both novel and known junctions. According to the authors' claim, SOAPsplice found more splice junctions than the other compared methods, and 97.24% of detected junctions were reported by more than one method. Huang et al. [2011] claim that although their method found fewer novel junctions than methods by Au et al. [2010] and Wang et al. [2010], but the percentage of junctions that are reported by more than one method for SOAPsplice (85.34%) is significantly higher than the other algorithms (TopHat: 67.73%, SpliceMap: 63.24%, MapSplice: 77.54%).

Huang et al. [2011] claim that their method is more efficient in detecting novel splice junctions as it outperforms all other algorithms with various read lengths and read depths specially when sequencing depth is lowest. This is very important considering that new junctions are usually found in low abundance parts of the transcript. The authors claim that their method is able to detect more genuine splice junctions than the compared methods.

As it is described by Zhang et al. [2012], RNA-Seq may be used for cellular phenotyping and help establishing the etiology of diseases characterized by abnormal splicing patterns. Recent studies have revealed that variations in splicing patterns are associated with Alzheimers and other complex diseases. In RNA-Seq, the exact nature of splicing events is buried in the reads that span exon-exon boundaries. The accurate and efficient mapping of these reads to the reference genome is a major challenge.

Zhang et al. [2012] refer to previous work by Trapnell et al. [2009], Dimon et al. [2010], and Wang et al. [2010].

The authors claim that the methods developed by Trapnell et al. [2009], Dimon et al. [2010] and Wang et al. [2010] do not have the ability to detect junctions without known splicing motifs. They state that both HMMSplicer by Dimon et al. [2010] and MapSplice by Wang et al. [2010] potentially work better for long reads than for short reads and they are less accurate on highly abundance transcripts. They also claim that neither of these two methods exploit the paired information in their algorithms.

PASSion does the splice junctions finding in five stages including initial mapping, building exon islands, high-resolution remapping, filtering and detection of canonical and non-canonical junctions. After initial mapping by a fast aligner, exon island are built by piling up the mapped reads. Pairs of one exonic read and one unmapped read are used as the basis of junction identification. These pairs are remapped, using pattern growth algorithm, to the reference genome and a splice junction is reported if the unique substrings from both ends can reconstruct the original split read and has a sufficiently high number of supportive reads.

Zhang et al. [2012] state that they analyzed their method on both simulated data and real data. They claim that they compared performance of PASSion with TopHat by Trapnell et al. [2009], MapSplice by Wang et al. [2010] and HMMSplicer by Dimon et al [2010] on these datasets.

The authors claim that on simulated data, their method, alongside other three tested methods, can detect almost all the true junctions when coverage is $> 100\times$fold. They note that PASSion predicted 136,664 and 172,568 splicing events for the two real datasets, of which 84.1% and 80.3% are known junctions.

Zhang et al. [2012] state that on the short read library of simulated data, the method by Trapnell et al. [2009] showed the least sensitivity comparing to other methods, and on libraries with long read, MapSplice by Wang et al [2010] detects the lowest number of junctions. The authors claim that in all simulated datasets, the true positive rate of PASSion has the quickest growth rate along with the read coverage and it is the most sensitive method overall. Zhang et al. state that when the specificity of TopHat, MapSplice and HMMSplicer drops with the read coverage, PASSion's specificity remains high with specificities of more than 97%.

The authors claim that for real datasets, in general, PASSion displays a balanced performance with both a high number of predictions and high confirmed ratios. They state that the pattern growth algorithm which is used in their approach, has not been taken advantage of in RNA-Seq analysis before. Zhang et al. [2012] note that PASSion can detect junctions with unknown motifs, which other three methods were unable to do so. They state that their method is the third fast method among other tested methods in terms of CPU hours.

3. CONCLUSION

3.1. Summary

This survey reviewed 10 journal papers published from late 2008 until early 2012. Half of the papers are published in the *Bioinformatics* journal of *Oxford Journals*, which is one of the most well-known journals in the field of bioinformatics. The other half were been published in more biologically-related journals, consisting of *Nucleic Acids Research* of *Oxford Journals*, *Genome Biology*, *Frontiers in Genetics* and *PloS ONE*.

The work of Trapnell et al. [2009] has been cited by 8 out of 10 papers that had been reviewed in this survey. This means that except the work of De Bona et al. [2008] that have been published before that, all consequent works on this subject referred to it. Furthermore, the method developed by Trapnell et al. [2009], was the only method that all of other methods used it as a basis for evaluating the performance of their own work.

Year	Author	Package Name	Notes
2009	Trapnell et al.	TopHat	Been cited by all papers, that reviewed in this survey, which published after it. The most cited paper in overall. The package is being updated very often and source code is freely available.
2010	Wang et al.	MapSplice	Cited by Huang et al. [2011], and Zhang et al. [2012] The package is being updated regularly and source code is freely available.
2011	Huang et al.	SOAPsplice	The package is being updated regularly and the package is freely available.
2012	Zhang et al.	PASSion	The latest published package reviewed in this survey. Source code is freely available.

3.2. Future Works

De Bona et al. [2008] state that their method can be extended to exploit homo-polymer errors, which is available for Roche's 454 sequencing platform.

Trapnell et al. [2009], authors of the most cited paper discussed in this survey, suggest that using paired-end read will drastically reduce the number of false positives in TopHat, and also improves its performance. Au et al. [2010], Lou et al. [2011a], and Zhang et al. [2012] developed their methods to exploit paired-end read information.

Huang et al. [2011] mention that in the future, their method could be optimized to run faster and consume less memory. We observed that SOAPSplice, the method presented by Huang et al. [2011], have been updated after publishing of the paper to reduce the amount of memory usage during generating the output.

In the latest published paper reviewed in this survey, Zhang et al. [2012] state that their method had missed some rare cross-chromosome splicing events, because it has been assumed that two reads map to the same chromosome. They suggested working to resolve this issue.

3.3. Acknowledgements

I would like to thank Dr Richard Frost and Dr Luis Rueda for their helpful comments and suggestions regarding this work.

Year	Author	Title of Paper	Major Contribution
2008	De Bona et al.	Optimal spliced alignments of short sequence reads.	**QPALMA**, one of the first works to address splice junction finding on RNA-Seq data. Use of SVM to find splice junctions.
2009	Trapnell et al.	**TopHat**: discovering splice junctions with RNA-Seq.	Introduces the concept of anchor as a way. Presents the idea of generating exon coverage islands.
2010	Au et al.	Detection of splice junctions from paired-end RNA-seq data by **SpliceMap**.	Designed to use information of paired-end reads. First method to use half-read mapping. Use of hash table for mapping.
2010	Ameur et al.	Global and unbiased detection of splice junctions from RNA-seq data.	**SplitSeek**, splits reads to two fragments and map them independently as anchors.
2010	Bryant et al.	**Supersplat**– spliced RNA-seq alignment.	Employs empirical RNA-Seq data for splice junction detection.
2010	Wang et al.	**MapSplice**: accurate mapping of RNA-seq reads for splice junction discovery.	Defining minimum anchor length as a filtering strategy. Comprehensive experiments on effect of various criteria including noise.
2010	Dimon et al.	**HMMSplicer**: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data.	Employs hidden Markov model to determine the exon boundaries.
2011	Huang et al.	**SOAPsplice**: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data.	Claims to achieve a better performance than other major methods using more Memory and more computing power.
2011	Lou et al.	Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length	Incorporates MLE method to align paired-end reads into reference genome. Introduces geometric-tail distribution for intron lengths.
2012	Zhang et al.	**PASSion**: A Pattern Growth Algorithm Based Pipeline for Splice Junction Detection in Paired-end RNA-Seq Data.	Introduces Pattern Growth algorithm to remap the reads. The ability to identify junctions with unknown splicing motifs.

4. ANNOTATIONS

4.1. Ameur et al. 2010

Citation. Ameur, A., Wetterbom, A., Feuk, L., and Gyllensten, U. 2010. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biology 11*, 3, R34.

Problem. High-throughput sequencing of mRNA opens extraordinary opportunities to identify the spectrum of splice events in a sample on a global scale. An important application of deep RNA sequencing is the discovery of fusion transcripts in cancer. Certain fusion transcripts are commonly produced by cancer cells, and detection of fusion transcripts is part of routine diagnostics of certain cancer types.

Previous Works. The authors refer to the work by Trapnell et. al [2009].

Shortcomings of Previous Work. Ameur et al. [2010] state that in the method designed by Trapnell et al. [2009], a substantial number of true splice junctions including junctions with long introns or non-canonical splice sites are outside of the detection range, and also this method is computationally challenging for transcripts expressed at lower levels.

New Idea/Algorithm/Architecture. The authors state that their method consists of a combination of a split-read alignment and the novel SplitSeek program. The alignment is performed using the AB/SOLiD whole-transcriptome-alignment software. The proposed method by Ameur et al. [2010], SplitSeek, developed in a way to find junction reads in which as few as five bases overlap with the other exon. It finds exon-exon boundaries that are supported by several split reads, it is required for each junction to be covered by at least two reads with unique starting points.

Experiments/Analysis Conducted. The authors state that they evaluated their method using public RNA-seq data from single mouse oocytes which was performed on two independent samples which consist of 50-bp reads.

Results. The authors state that they selected 22 base pair for the anchor length according to the highest number of uniquely mapped split reads was obtained for this length. Ameur et al. present the results of their experiment in terms of number of splice junctions and insertions, number of predicted small insertions and deletions within RefSeq exons, and number of predicted splice junctions as a function of the total number of processed reads.

Claims. The authors claim that the exon-exon boundaries are identified almost at nucleotide resolution and with a low false-positive rate, less than one in 10,000, for junctions within 100 kb. Ameur et al. state that their method makes it possible to study splice junctions and fusion genes while also quantifying the gene expression using RNA-Seq data. They claim according to their results, that their proposed algorithm has a very low false-positive rate, they state that acquired false discovery rate of less than one for 1,000 junctions within 1Mb and less than 1/10,000 for those within 100 kb, is supporting their claim.

Citations By Others. Dimon et al. [2010] and Lou et al. [2011a].

4.2. Au et al. 2010

Citation. 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 Research 38*, 14, 4570-8.

Problem. Abnormal RNA splicing is associated with many human diseases. For this reason, methods to identify and quantify splicing events are important to biology and medicine.

Previous Works. The authors refer to the work by Mortazavi et al. [2008] and Trapnell et al. [2009].

Shortcomings of Previous Work. The authors state that the method developed by Mortazavi et al. [2008] is dependent on annotated exon library and since the exon library is incomplete, this method can not find junctions that involve novel splicing events. The authors do not mention any shortcomings of the method presented by Trapnell et al. [2009].

New Idea/Algorithm/Architecture. Au et al. [2009] present their method, SpliceMap, based on the idea using the mapping of half-reads as a way to identify the approximate location of a junction. SpliceMap works in four steps, half-read mapping, seeding selection, junction search and paired-end filtering. It maps both halves of the read to the reference genome by a short read mapping tool.

Experiments/Analysis Conducted. The authors state that they compared their method with the method described by Trapnell et al. [2009] on an RNA-Seq dataset of 23,41,226 reads. They claimed that they assessed their method's specificity by aligning detected junctions to human ESTs in GenBank. They also stated that they investigated novelty of discovered junctions by PCR experiments. Au et al. [2010] describe a comparison with ERANGE method by Mortazavi et al. [2008]. They also stated that they compared their method with BLAT, which is common tool for EST sequences alignment. The authors claimed that they calculated the performance of their method in a specific CPU running time and compared it with TopHat by Trapnell et al. [2009].

Results. Au et al. [2010] claim that 87.9% of junctions found by their method were supported by EST evidence. They state that SpliceMap achieves more than 95% sensitivity for highly expressed genes, more than 90% for genes with medium expression and (40-67%) for genes with low expression. They state that more genes detected by SpliceMap are of higher degree (80100%) of completeness in junction discovery. Au et al. [2010] claim that in a random sample experiment, 85% of novel junctions validated using PCR experiment.

The authors stated that ERANGE package by Mortazavi et al. [2009] found 160,899 junctions and SpliceMap found 151,317 junctions, among those found by [Au et al. 2010] method, 23,020 junction, which were not found by ERANGE, were novel. They also claimed that the BLAT package, achieved a similar but still slightly lower level of specificity with a much lower sensitivity (70% lower) as compared to SpliceMap.

The authors stated that it took 66 CPU hours for SpliceMap and 12 CPU hours for ERANGE to process the data set.

Claims. They claim that based on their results, SpliceMap detects more annotated junctions than TopHat method by Trapnell et al. [2009]. They claim that 50bp reads can support an approach of direct de novo detection of splice junctions without the need to first cluster reads to identify accepted exons, and that this approach can achieve significantly higher sensitivity in junction detection than current leading methods of RNA-seq analysis. They also claim that paired-read information can help to reduce false discoveries.

Citations By Others. Wang et al. [2010], Dimon et al. [2010], Lou et al. [2011a], and Huang et al. [2011].

4.3. Bryant et al. 2010

Citation. Bryant, D. W., Shen, R., Priest, H. D., Wong, W.-K., and Mockler, T. C. 2010. Supersplat-spliced RNA-seq alignment. *Bioinformatics (Oxford, England) 26*, 12,

1500–5.

Problem. Next-generation sequencing provides a ground for massive transcript expression analysis. RNA-Seq approach supplies enough reads for this purpose, and the key to profiling this genetic information is identification of intron/exon boundaries or splice junctions.

Previous Works. The authors refer to the work by De Bona et al. [2008], Trapnell et al. [2009], and Filichkin et al. [2010].

Shortcomings of Previous Work. Bryant et al. [2010] state that the method developed by De Bona et al. [2008] relies on the previously known splice sites for training the algorithm which influences the results. Also they note that QPALMA scores junctions that conforms with canonical splicing motifs higher, so it may be inefficient in finding non-canonical splice junctions. The authors state that this problem also applies to TopHat, the method developed by Trapnell et al. [2009]. Bryant et al. [2010] state that TopHat needs a high number of RNA-Seq reads to build exon islands.

New Idea/Algorithm/Architecture. Bryant et al. [2010] introduce a new approach, called Supersplat, that uses hash table as a way to save system memory by storing read sequences as key and their frequencies as the value. Supersplat uses two parameters to limit the maximum and minimum length of the sequence which it is trying to build location indexes based upon them. After indexing the reference sequence, Supersplat identifies reads that can be aligned against the reference, as possible splice junctions in an iterative process. Potential splice junctions are filtered according to the number of overlapping reads on two intron boundaries.

Experiments/Analysis Conducted. The authors state that they tested the performance of their method on a set of 3,690,882 *Arabidopsis thaliana* reads. They used TAIR8 database of annotated junctions to evaluate Supersplat's performance. Bryant et al. [2010] also state that they assessed their approach for *de novo* splice junction discovery on a dataset of *Brachypodium distachyon*.

Results. The authors claim that they confirmed 91% of canonical and 86% non-canonical splice junctions using PCR and Sanger sequencing. Bryant et al. [2010] claimed that they achieved a predicted positive rate (PPV) of 70% with the minimum read length of 6 and a 90% rate by increasing it. According to Bryant et al. [2010], this rate reach 97% by setting the overlapping number of reads filter to 21.

Claims. Bryant et al. [2010] claim that their approach is unbiased and exhaustive but it may generate output files with up to tens of gigabytes size, and the user should account for determination of the befitting criterion to filter out spurious output. The authors claim that the exhaustive approach of their method has the ability to discover many previously unknown splice junctions.

Citations By Others. Dimon et al. [2010], Huang et al. [2011], and Lou et al. [2011a].

4.4. De Bona et al. 2008

Citation. De Bona, F., Ossowski, S., Schneeberger, K., and Rätsch, G. 2008. *Optimal spliced alignments of short sequence reads. Bioinformatics (Oxford, England) 24*, 16, i174–80.

Problem. Short reads acquired from high-throughput sequencing technologies can be used for studying transcriptome and gene structure identification. Aligning these reads over intron/exon boundaries is a requirement for this purpose.

The authors do not specify the reason for studying transcriptome explicitly.

Previous Works. The authors do not refer to any previous work which relates to the subject of this survey.

Shortcomings of Previous Work. The authors do not refer to any previous work.

New Idea/Algorithm/Architecture. De Bona et al. [2008] developed their approach, called QPALMA, in three independent parts, splice site prediction model, dynamic programming algorithm and scoring function. QPALMA aims to align short reads to the reference genome, splice site prediction is helping this approach to achieve better results. This part is based on a machine learning approach which uses a set of donor and acceptor sites to train a SVM predictor. The authors propose three different extensions for Smith-Waterman algorithm for aligning the reads to the reference.

Experiments/Analysis Conducted. The authors state that they trained their algorithm using a simulated dataset of 10,000 previously aligned sequences. The alignment error rate for these rates have been calculated incorporating different available informations. The authors also tried their approach on a dataset of spliced and unspliced 2.98 million reads of forward strands of chromosome 1.

Results. De Bona et al. [2008] claim that they could align 10,000 *in silico* spliced reads with an error rate of 1.78% incorporating quality information, intron length model and splice site predictions. The authors state that it was the best rate that they have achieved. The authors also claim that QPALMA aligned spliced and unspliced reads with a 5.2% and 1.2% error rate respectively.

Claims. The authors claim that they could successfully exploit all information sources to align short reads over exon boundaries. The authors claim that their approach works reasonably for all next-generation sequencing platforms including *Illumina* sequencing which have been tried in their experiment.

The authors state that their method can be extended to exploit homo-polymer errors which is available for Roche's 454 sequencing platform.

Citations By Others. Trapnell et al. [2009], Bryant et al. [2010], Wang et al. [2010], and Huang et al. [2011].

4.5. Dimon et al. 2010

Citation. 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*, 11, e13875.

Problem. During the past decade, there has been a growing appreciation of the importance of alternative splicing as a mechanism for organisms to increase proteomic diversity and regulatory complexity. According to authors, the ability to detect alternative splice isoforms with accuracy and sensitivity is key to comprehensive RNA-Seq analysis.

Previous Works. The authors refer to previous work by Mortazavi et al. [2008], Trapnell et al. [2009], Bryant et al. [2010], and Ameur et al. [2010].

Shortcomings of Previous Work. The authors note that the method developed by Mortazavi et al. [2008] does not address the question of novel junctions and cannot be used for organisms with incomplete or inaccurate genome annotations. They state that the algorithm developed by Trapnell et al. [2009] performs best on mammalian transcripts with relatively high abundance, but has defects in more compact genomes and with non-canonical junctions. They note that the method proposed by Ameur et al. [2010] has the requirement for at least one read to split evenly across the exon-exon boundary which reduces sensitivity in low coverage datasets and transcripts, also they claim that this method supports only ABI SOLiD reads.

Dimon et al. [2010] state that SuperSplat, the method developed by Bryant et al. [2010], requires both pieces of a read to be exact matches to the reference sequence and conclude that it is not robust against sequencing errors or single-nucleotide polymorphisms. The authors claim that the algorithm designed by Au et al. [2010] considers only canonical splice junctions and requires read lengths of 50bp or greater.

New Idea/Algorithm/Architecture. Dimon et al. [2010] claim that they have developed a method to avoid the inherent bias introduced by relying upon previously defined biological information. Their algorithm, called HMMSPlicer, works by dividing each read in half and seeding the read-halves against the genome and using a Hidden Markov Model to determine the exon boundary. They claim that both canonical and non-canonical junctions are reported and a score is assigned to each junction, dependent on the strength of the alignment and the number and quality of bases supporting the splice junction.

Experiments/Analysis Conducted. The authors claim that they compared their algorithm with algorithms designed by Trapnell et al. [2009] and Au et al. [2010]. They state that they analyzed the performance of their method on simulated reads and three publicly available experimental datasets.

Results. Dimon et al. include the detailed results of their experiments with different algorithm parameters on tested datasets. The authors state that in comparison with TopHat, HMMSplicer shows its ability to find more junctions with a similar level of specificity in each of tested datasets. They state that in comparison with SpliceMap by Au et al. [2010], their method achieve 7% more matching junctions for human datasets and it outperforms SpliceMap in the low sequence quality *A. thaliana* dataset.

Claims. The authors claim that HMMSplicer combines high sensitivity with a low false positive rate, performs well on datasets with low quality sequence reads, performs well in datasets with uneven coverage, identifies many junctions in low abundance transcripts and also identifies non-canonical junctions, and finds true novel junctions in genomes with incomplete annotation.

The authors claim that their algorithm is the only software package that provides a score for each junction, reflecting the strength of the junction prediction.

Citations By Others. Huang et al. [2011] and Zhang et al. [2012].

4.6. Huang et al. 2011

Citation. 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*, July, 1–12.

Problem. Splice junction detection is the first step of studying alternative splicing. Alternative splicing is highly effective on diversity of proteins, as it causes different mRNAs to be produced from the same gene. These different mRNAs translate into different protein isoforms.

Previous Works. The authors refer to the previous work by Mortazavi et al. [2008], De Bona et al. [2008], Trapnell et al. [2009], Bryant et al. [2010], Au et al. [2010], Wang et. al [2010], and Dimon et al. [2010].

Shortcomings of Previous Work. The authors state that QPALMA, the method developed by De Bona et al. [2008] which uses a machine learning approach, is biased toward splice junctions that are similar to the ones in the training data set. Huang et al. [2011] state that low sequencing depth affects the performance of the algorithm developed by Trapnell et al. [2009] and hence there would not be enough reads for efficient junction detection. The authors state that the method introduced by Bryant et al. [2010], which uses hashing as its alignment approach, needs a large amount of memory and computing power and as a result is not scalable for reads longer that 50 base pairs.

Huang et al. [2011] state that SpliceMap, the algorithm presented by Au et al. [2010], performs poorly while dealing with the reads that can be mapped to more than one location. Also they state that this approach is not efficient when the transcriptome is lowly expressed and the reads have sequencing errors. The authors state that the method developed by Wang et al. [2010] has some inefficiencies while the sequencing depth is low and leads to reduced call rate.

New Idea/Algorithm/Architecture. Huang et al. [2011] present SOAPsplice, which finds the splice junctions in two steps. In the first step, it maps the reads into reference genome using Burrows Wheeler Transformation for indexing and detects splice junction candidates based on some criteria which includes following known splicing motifs and a maximum intron size of 50,000 bp. SOAPsplice applies two different filtering techniques to omit false positives, the first strategy is to check the paired-end information with the direction of mate-pair reads and discard incompatible junctions. The other strategy is to filter out the junctions that have a missing segment between two sub-reads that have been mapped to the reference genome.

Experiments/Analysis Conducted. Huang et al. [2011] compare their method with the algorithms developed by Trapnell et al. [2009], Wang et al. [2010], and Au et al. [2010] on two 50 and 150 bp simulated dataset and two 51 and 130 bp real dataset.

Results. The authors claim that according to the results of the simulated dataset for both 50 and 150 bp length reads, their method had the highest call rate while it kept the false positive rate at its lowest comparing to other approaches. For the real dataset with 51 bp reads, Huang et al. [2011] claim that SOAPsplice detect more novel junctions than TopHat by Trapnell et al. [2009] and its results are comparable with the method designed by Au et al. [2010] on both novel and known junctions. According to the authors' claim, SOAPsplice found more splice junctions than the other compared methods, and 97.24% of detected junctions were reported by more than one method. Huang et al. [2011] claim that although their method found less novel junctions than methods by Au et al. [2010] and Wang et al. [2010], but the percentage of junctions that are reported by more than one method for SOAPsplice (85.34%) is significantly higher than the other algorithms

(TopHat: 67.73%, SpliceMap: 63.24%, MapSplice: 77.54%).

Claims. Huang et al. [2011] claim that their method is more efficient in detecting novel splice junctions as it outperforms all other algorithms with various read lengths and read depths specially when sequencing depth is lowest. This is very important considering that new junctions are usually found in low abundance parts of the transcript. The authors claim that their method is able to detect more genuine splice junctions than the compared methods.

Citations By Others. None.

4.7. Lou et al. 2011

Citation. Lou, S.-K., Li, J.-W., Qin, H., Yim, A., Lo, L.-Y., Ni, B., Leung, K.-S., Tsui, S., and Chan, T.-F. 2011. Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length. *BMC Bioinformatics 12*, Suppl 5, S2.

Problem. Studying the way that alternative splicing affects a biological system is as important as studying its fundamental regulatory mechanisms, and as RNA-Seq provides the ability to analyze transcriptome in a base-level resolution and high coverage.

Previous Works. The authors refer to the work by Mortazavi et al. [2008], Trapnell et al. [2009], Ameur et al. [2010], Au et al. [2010], Bryant et al. [2010], and Lou et al. [2011b].

Shortcomings of Previous Work. Lou et al. [2011a] state that the approach presented by Trapnell et al. [2009] depends on canonical splice site motifs. The authors also state that as the methods developed by Ameur et al. [2010], Au et al. [2010] and Bryant et al. [2010] designed based on the idea of read counting, sequencing depth can significantly affect their performance.

New Idea/Algorithm/Architecture. The authors proposed an approach based on maximum likelihood estimation which relies on geometric-tail distribution of intron lengths for aligning of paired-end RNA-Seq reads. The authors used a package named ABMapper which was particularly developed for the purpose of spliced mapping by the same team as the authors and is explained in [Lou et al. 2011b]. They state that their approach is an empirical probabilistic model which adopted a two part distribution, an arbitrary length distribution and a geometric distribution. This method use maximum likelihood to estimate the most probable location for a paired-end read based on this two-part distribution. The authors stated that their approach works in three models, one without any *a priori* knowledge, and two with expression level and junction-site frequency as *a priori* knowledge.

Experiments/Analysis Conducted. The authors state that they compared their model with methods developed by Trapnell et al. [2009] and Au et al. [2010]. They used two human lymphoblastoid cell-line datasets for testing purpose. The dataset consist of 8.4 million 75 bp paired-end reads with an approximately 250 bp insert size. The results were validated with Alternative Splicing and Transcript Diversity (ASTD) database and Human EST database.

Results. Lou et al. [2011a] claim that their method was able to report 53% and 49% more splice junctions comparing to the methods by Au et al. [2010] and Trapnell et al. [2009]. The authors claim that 60% of the junctions that was predicted only by their method

could be validated by ASTD which comprised 22% of the total reported splice junctions. According to the authors, this implies that the methods proposed by Au et al. [2010] and Trapnell et al. [2009] missed at least one-fifth of the true splice junctions. The authors claim that by performing an exhaustive search for junctions in Human EST database, they found that their method predicted splice junctions with an accuracy of 96%.

Claims. The authors claim that their proposed approach is able to detect 50% more splice junctions than other existing tools. Lou et al. [2011a] claimed that the reason for superiority of their approach is in using first, ABMapper which has a much higher sensitivity in spliced-mapping than other approaches, and the second is the geometric-tail based model.

Citations By Others. None.

4.8. Trapnell et al. 2009

Citation. Trapnell, C., Pachter, L., and Salzberg, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England) 25*, 9, 1105-11.

Problem. Alternative splicing is an important process in normal cellular functions and also in human diseases. Finding novel splice junctions is an important part of studying alternative splicing.

Previous Work. The authors refer to previous work by De Bona et al. [2008] and Mortazavi et al. [2008].

Shortcomings of Previous Work. The authors state two shortcomings of De Bona et al's work, the first is that their method, QPALMA, depends on a set of known splice junctions from the reference genome and cannot identify novel ones. They state that the other shortcoming is that De Bona et al. use Vmatch, an alignment program which is not designed to map short reads on machines with small main memories and is considerably slower than other short-read mappers. Trapnell et al. [2009] state that ERANGE, the method developed by Mortazavi et. al, depends on available annotation of exon-exon junctions for its main objective which is gene expression quantification in mammalian RNA-Seq projects.

New Idea/Algorithm/Architecture. The authors introduce a new system called TopHat which works in two phases to find junctions. In the first phase all reads are mapped to the reference genome using Bowtie, all reads that do not map to the reference genome are set aside as initially unmapped reads. Then an initial consensus of mapped regions, called exon islands, is computed using the assembly module in a package named Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. For each splice junction, Tophat searches the initially unmapped reads in order to find reads that span junctions using a seed-and-extend strategy.

Experiments/Analysis Conducted. Trapnell et al. [2009] state that they conducted an experiment on 47,781,892 short reads using their method, TopHat, and a previously developed method called ERANGE by Mortazavi et al. [2008].

Results. The authors claim that their method was able to discover around 72% of splice junctions comparing to annotation-based analysis done in [Mortazavi et al. 2008] in less transcribed regions and 80% of junctions in more actively transcribed regions. They claimed that out of 19,722 newly discovered junctions that they found in their experiment, many of them are true splices but it is difficult to asses exactly how many of them are

genuine.

Claims. The authors claim that the significance of their work is in its ability to detect novel splice junctions. They also claimed that their tool represents a significant advance over previous RNA-Seq splice detection methods.

Citations By Others. Au et al. [2010], Bryant et al. [2010] , Wang et al. [2010], Ameur et al. [2010], Dimon et al. [2010], Huang et al. [2011], Lou et al. [2011a], and Zhang et al. [2012].

4.9. Wang et al. 2010

Citation. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. a., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. 2010. Map-Splice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research 38*, 18, e178.

Problem. Accurate identification and quantification of transcript isoforms is crucial to characterize alternative splicing among different cell types. In addition, sequence variants found within splice sites or splicing enhancer sequences may have functional consequences on alternative splicing. A large proportion of human genetic disorders result from splicing variants.

Previous Works. The authors refer to the work by De Bona et al. [2008], Trapnell et al. [2009] and Au et al. [2010].

Shortcomings of Previous Work. The authors note that the output generated by the method of Au et al. [2010] does not include tag alignments, hence is incomplete. They do not state any shortcoming regarding the works of others.

New Idea/Algorithm/Architecture. According to Wang et al. [2010], their method operates in two phases. In the first phase, called tag alignment, candidate alignments of the mRNA tags to the reference genome are determined. A set of candidate alignments are computed for each tag as multiple possible alignments may be found for each read. In the second phase which is called splice inference phase, splice junctions that appear in the alignments of one or more tags are analyzed to determine a splice significance score based on the quality and diversity of alignments that include the splice. The most likely alignment for each tag is chosen based on the splice significance score.

Experiments/Analysis Conducted. The authors state that they evaluated specificity and sensitivity of their method using an experiment on a generated synthetic dataset. They also state that they validated their method using qRt-PCR experiment.

Results. [Wang et al. 2010] state that they achieved a true-positive rate of 96% and false-positive rate of 8% for their method. They stated that over 77% of canonical junctions found by their method were confirmed by known transcripts in GenBank, which was between 6% to 11% higher in comparison by TopHat method by Trapnell et al. [2009],

Claims. The authors claim that both TopHat by Trapnell et al. [2009] and their method were more memory efficient and much faster in experiments than SpliceMap by Au et al. [2010]. They also claim that their method performed best by detecting more true-positive junctions and fewer false-positive junctions than the other two methods. They state that longer tags improve both the sensitivity and the specificity of the junction discovery in their

method and also in method by Trapnell et al. [2009], and they claim that in comparison, their method has a higher sensitivity in different tag lengths. They claim that using read lengths of 75 or 100bp yield to significantly better sensitivity and specificity for splice detection.

Citations By Others. Huang et al. [2011] and Zhang et al. [2012].

4.10. Zhang et al. 2012

Citation. Zhang, Y., Lameijer, E.-W.,'t Hoen, P. a. C., 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 (Oxford, England), 1-8.*

Problem. RNA-Seq may be used for cellular phenotyping and help establishing the etiology of diseases characterized by abnormal splicing patterns. Recent studies have revealed that variations in splicing patterns are associated with Alzheimers and other complex diseases. In RNA-Seq, the exact nature of splicing events is buried in the reads that span exon-exon boundaries. The accurate and efficient mapping of these reads to the reference genome is a major challenge.

Previous Works. The authors refer to previous work by Trapnell et al. [2009], Dimon et al. [2010], and Wang et al. [2010].

Shortcomings of Previous Work. The authors claim that the methods developed by Trapnell et al. [2009], Dimon et al. [2010] and Wang et al. [2010] do not have the ability to detect junctions without known splicing motifs. They state that both HMMSplicer by Dimon et al. [2010] and MapSplice by Wang et al. [2010] potentially work better for long reads than for short reads and they are less accurate on highly abundance transcripts. They also claim that neither of these two methods exploit the paired information in their algorithms.

New Idea/Algorithm/Architecture. PASSion does the splice junctions finding in five stages including initial mapping, building exon islands, high-resolution remapping, filtering and detection of canonical and non-canonical junctions. After initial mapping by a fast aligner, exon island are built by piling up the mapped reads. Pairs of one exonic read and one unmapped read are used as the basis of junction identification. These pairs are remapped, using pattern growth algorithm, to the reference genome and a splice junction is reported if the unique substrings from both ends can reconstruct the original split read and has a sufficiently high number of supportive reads.

Experiments/Analysis Conducted. Zhang et al state that they analyzed their method on both simulated data and real data. They claim that they compared performance of PASSion with TopHat by Trapnell et al. [2009], MapSplice by Wang et al. [2010] and HMMSplicer by Dimon et al [2010] on these datasets.

Results. The authors claim that on simulated data, their method, alongside other three tested methods, can detect almost all the true junctions when coverage is $> 100\times$fold. They note that PASSion predicted 136,664 and 172,568 splicing events for the two real datasets, of which 84.1% and 80.3% are known junctions.

Claims. Zhang et al. [2012] state that on the short read library of simulated data, the method by Trapnell et al. [2009] showed the least sensitivity comparing to other methods, and on libraries with long read, MapSplice by Wang et al [2010] detects the lowest number

of junctions. The authors claim that in all simulated datasets, the true positive rate of PASSion has the quickest growth rate along with the read coverage and it is the most sensitive method overall. Zhang et al. state that when the specificity of TopHat, MapSplice and HMMSplicer drops with the read coverage, PASSion's specificity remains high with specificities of more than 97%.

The authors claim that for real datasets, in general, PASSion displays a balanced performance with both a high number of predictions and high confirmed ratios. They state that their method is using pattern growth algorithm, which has not been used in RNA-Seq analysis before. Zhang et al. [2012] note that PASSion has the ability to detect junctions with unknown motifs, which other three methods was unable to do so. They state that their method is the third fast method among other tested methods in terms of CPU hours.

Citations By Others. None

REFERENCES

Ameur, A., Wetterbom, A., Feuk, L., and Gyllensten, U. 2010. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biology 11,* 3, R34.

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 Research 38,* 14, 4570–8.

Bryant, D. W., Shen, R., Priest, H. D., Wong, W.-K., and Mockler, T. C. 2010. Supersplat–spliced RNA-seq alignment. *Bioinformatics (Oxford, England) 26,* 12, 1500–5.

De Bona, F., Ossowski, S., Schneeberger, K., and Rätsch, G. 2008. Optimal spliced alignments of short sequence reads. *Bioinformatics (Oxford, England) 24,* 16, i174–80.

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,* 11, e13875.

Filichkin, S. a., Priest, H. D., Givan, S. a., Shen, R., Bryant, D. W., Fox, S. E., Wong, W.-K., and Mockler, T. C. 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research 20,* 1, 45–58.

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,* July, 1–12.

Lou, S.-K., Li, J.-W., Qin, H., Yim, A., Lo, L.-Y., Ni, B., Leung, K.-S., Tsui, S., and Chan, T.-F. 2011a. Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length. *BMC Bioinformatics 12,* Suppl 5, S2.

Lou, S.-K., Ni, B., Lo, L.-Y., Tsui, S. K.-W., Chan, T.-F., and Leung, K.-S. 2011b. ABMapper: a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics (Oxford, England) 27,* 3, 421–2.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods 5,* 7, 621–8.

Trapnell, C., Pachter, L., and Salzberg, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England) 25,* 9, 1105–11.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. a., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. 2010. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research 38,* 18, e178.

Zhang, Y., Lameijer, E.-W., 't Hoen, P. a. C., 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 (Oxford, England)*, 1–8.