Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 831352, 9 pages
http://dx.doi.org/10.1155/2015/831352

Review Article

Survey of Programs Used to Detect Alternative Splicing Isoforms from Deep Sequencing Data *In Silico*

Feng Min, Sumei Wang, and Li Zhang

Department of Infectious Diseases, The Affiliated Chenggong Hospital of Xiamen University, The 174th Hospital of the Chinese People's Liberation Army, Xiamen, Fujian 361000, China

Correspondence should be addressed to Sumei Wang; wangsumei174@126.com

Received 26 November 2014; Revised 17 February 2015; Accepted 2 March 2015

Academic Editor: Yunlong Liu

Next-generation sequencing techniques have been rapidly emerging. However, the massive sequencing reads hide a great deal of unknown important information. Advances have enabled researchers to discover alternative splicing (AS) sites and isoforms using computational approaches instead of molecular experiments. Given the importance of AS for gene expression and protein diversity in eukaryotes, detecting alternative splicing and isoforms represents a hot topic in systems biology and epigenetics research. The computational methods applied to AS prediction have improved since the emergence of next-generation sequencing. In this study, we introduce state-of-the-art research on AS and then compare the research methods and software tools available for AS based on next-generation sequencing reads. Finally, we discuss the prospects of computational methods related to AS.

1. Introduction

Alternative splicing (AS) refers to the production of pre-mRNA via gene transcription to generate a number of mature mRNAs based on different splice modes, thereby increasing protein diversity. Since alternative splicing was discovered, studies have identified a large number of AS events in the human gene transcription process [1]. Based on high-throughput deep sequencing data, AS occurs in approximately 95% of the human genome [2]. AS is an important regulatory mechanism involved in the regulation of eukaryotic gene expression and proteome diversity [3]. The process is closely linked with many diseases, including cancer and diseases of the nervous system [4–6]. Thus, scholars in medicine, genetics, bioinformatics, and other fields have directed considerable research interest towards AS with the aim of identifying additional splicing events that could facilitate a deeper understanding of the AS regulatory mechanism.

Splice site recognition represents a key step in selective splicing research. Splice sites are used to predict the positions of exon/intron structures and splice site features, and splice site recognition is the traditional strategy used to predict alternative splice sites. Many algorithms, software, and databases for sequence alignment have emerged due to the application of first-generation sequencing. The research resources designed specifically for AS have gradually become richer, including a common ASD AS database [7]. However, the cost of first-generation sequencing is high; considerable efforts have been directed towards the goal of creating thousand- and hundred-dollar genome sequencing technology in the postgenomic era. Thus, the high throughput and low cost of next-generation sequencing technologies have provided a new stage for scientific research [8, 9].

AS was discovered in 1977 [10]. Subsequently, researchers realized the importance of AS due to its ability to regulate gene expression and facilitate protein diversity [11, 12]. The advantages of next-generation sequencing technology have opened a new stage of sequencing, and the study of the massive amounts of data generated by RNA-seq technology has become an important research direction.

RNA-seq (high-throughput RNA sequencing) represents a new method for the analysis of gene expression and transcriptomes. Many software tools and databases have appeared with the capacity to generate short sequence alignments and predictions on the basis of the alternative splice sites identified using RNA-seq.

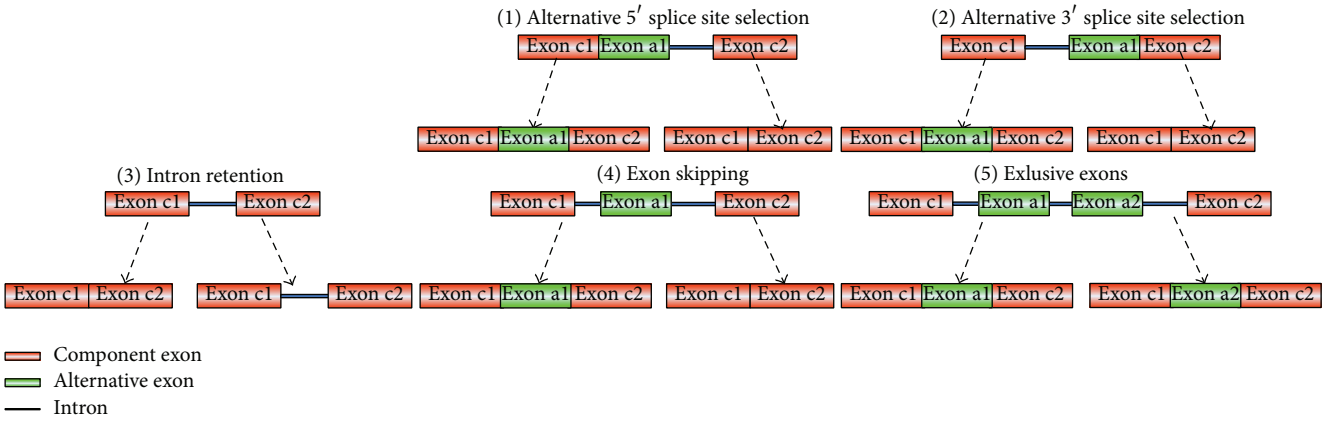

Figure 1: Five types of alternative splicing.

In this study, we outlined the methods, software tools, and databases available for AS research under two-generation sequencing technologies. The effect of these factors on AS research was analyzed. Using RNA-seq data produced by the Illumina/Solexa sequencing platform as an example, we compared three common splice site prediction programs (HMMSplicer [11], SOAPsplice, and TopHat [8]) under conditions of different depths and sequence read lengths. The performance of each type of software was evaluated under different conditions by comparing the number of accurately predicted sites, the accuracy rate, and the error rate. Finally, we discussed the problems and challenges associated with using deep sequencing data to study AS.

2. Discovering Alternative Splicing Sites from Long DNA Sequences

In addition to experimental methods, researchers predict potential AS events through the comparison between EST expression sequence tags and gene sequences. A large number of analyses and studies have validated the significance of the 3′ terminal splice acceptor site and 5′ terminal splice donor site in splicing events. Figure 1 summarizes the five AS forms.

The study by Fairbrother et al. [13] on exons in the human genome revealed that the splicing enhancers ESE and ESS serve an important regulatory function in selective splicing. Black [14] demonstrated that the splicing enhancer ISE and silencer ISS are also important for the selection of splicing sites and recognition of exons and introns. Thus, the AS process in eukaryotic genes is determined not only by a splicing factor but also by a complex regulatory process.

The means of selective splicing mainly include the following.

(1) Comparison analysis based on ESTs, mRNA, and gene fragments: EST comparative analysis was one of the earliest AS research methods. This method can identify certain AS events. However, EST has its own limitations, such as incomplete data, influence from genetic pollution, sensitive 3′ terminal, and high cost [15, 16]. Common comparison software programs include BLAT [17], Clustal [18], SIM4 [19], Ecgene [20], ASPIC [21], Spidey [22], GeneSeqer [23], and GMAP [24].

(2) Using gene chip high-throughput technology: gene chip technology has facilitated the research upsurge in the whole gene transcriptome. A large number of AS events have been identified using this technology. Johnson et al. [1, 25] discovered many exon-skipping events by analyzing microarray data. However, the disadvantage of this method is that probe density is limited, and designing a probe based on the known sequence and data analysis is difficult.

(3) Using machine learning methods for theoretical prediction: machine learning techniques have been widely used in various tasks in the field of bioinformatics, such as protein remote homology detection [26–29], microRNA identification [30, 31], protein binding site prediction [32], domain boundary identification [33, 34], DNA-binding protein prediction [35–37], protein structure prediction [38], enzyme classification [39, 40], gene regulation network construction [41], heat shock protein classification [42, 43], replication origin prediction [44, 45], nucleosome positioning sequence identification [46–48], CpG island methylation status prediction [49], translation initiation site prediction [50], promoter prediction [51], and microarray clustering [52, 53]. These machine learning based methods have achieved promising predictive performances. Therefore, some researchers have also applied common machine learning methods for theoretical predictions, such as support vector machine (SVM) [54, 55], weight matrices, the hidden Markov model, the quadratic discriminant function [56], and the neural network model [57]. The programs used for predicting splice sites based on these algorithms include HMMgene [58], NetGene2 [59, 60], geneID [61], GeneSplicer [62], and SpliceMachine [63].

3. Discovering Alternative Splicing Sites from Short Reads

The next-generation high-throughput sequencing technology developed rapidly after its emergence, thus enabling sequencing technology to move a step closer towards the thousand-dollar genome project. RNA-seq represents a new approach for gene expression and transcriptome studies. Currently, traditional AS research methods coexist with the development of the next-generation research methods. An increasing number of studies have been devoted to the development of new algorithms. In summary, next-generation high-throughput sequencing technology can provide a broad platform for AS due to its high efficiency and inexpensiveness.

However, RNA-seq also has shortcomings. The main challenge stems from read length. The read length of first-generation sequencing (i.e., Sanger sequencing) reaches approximately 1000 bp. The initial read length of RNA-seq was only approximately 25 bp. The read length is still relatively short, despite reaching 100 bp using Illumina/Solexa double-end sequencing [64].

3.1. Data Preprocessing. The first step in predicting an alternative splice site is to position the read on the reference transcriptome using RNA-seq data. However, the general analysis tools often position the reads on the reference genome because the transcriptome itself is not complete [8]. Short RNA-seq read lengths and incomplete transcriptomes cause the accuracy of this step to directly influence the accuracy of the prediction.

Some data found in read mapping can cross two exon-exon junctions [65]. This "read in junction" cannot be directly positioned on the genome sequence. This finding represents the key to studying alternative splice sites and identifying the critical region for exploring undetected splice events. Therefore, the processing strategy used to splice the read in junction is the key to predicting splice sites [66]. One approach for the treatment of read in junction is to position the reads onto the reference genome according to the currently known annotation of the exons. ERANGE [67] uses this method. Obviously, identifying new splice events is difficult using this approach. Another approach is to completely position the reads on the reference genome so that they can be divided into several different clusters. Reads with overlapping areas are classified into the same cluster. An exon region is delimited in each cluster [65]. Finally, the reads in junctions are positioned on the possible junctions. New splice events can be identified because the reads are based on known exon annotations. The splice site prediction software TopHat [8] uses this strategy.

Numerous software programs are specifically designed for the read mapping of RNA-seq data. These programs adopt the following algorithms: (1) the Smith-Waterman algorithm, such as BFAST [68] and SHRiMP [69]; (2) the two-way Burrows-Wheeler transform (BWT) algorithm, such as SOAPAligner [70]; (3) the BWT algorithm, such as Bowtie [71] and BWA [72]; and (4) the spaced-seed vacancy seed algorithm, such as MAQ [73]. Data compatibility should also be considered along with the choice of software. The formats of RNA-seq data generated by various sequencing platforms are different [74]. Thus, software versatility is affected by the styles and variety of formats it supports. Bowtie and BWA are relatively efficient, whereas SOAPAligner, BFAST, and MAQ have good tolerance for mismatches.

In addition to read mapping, we identified special software devoted to read assembly (i.e., de novo assembly). Few methods to study AS based on read assembly exist. However, read assembly has special roles in other biological information sciences. The typical read assembly software includes SHARCGS [75], SSAKE [76], and ALLPATHS [77]. The former two are assembled only for single sequence data, while the latter can be assembled for a pair of sequences from double-end sequencing. MAQ also has the ability to perform read assembly. Finally, sequence read archive (SRA) files are specialized for the storage of databases related to RNA-seq data for NCBI for inclusion into an AS database.

3.2. Alternative Splicing Prediction. The common AS site prediction software includes ERANGE, QPALMA [78], TopHat, MapSplice [79], SpliceMap, SOAPsplice, SplitSeek [80], and HMMSplicer. Current studies using RNA-seq to identify AS sites focus on locating splice sites, discovering new splice sites located as distantly as possible, and conducting next-step AS studies. Therefore, the accuracy and efficiency of predictions are key factors for the prediction software. Moreover, accuracy should be improved in order to predict more splice sites, while the error probability should be reduced; these factors differ for selected algorithms.

ERANGE was the earliest available method. It was the first program to use the read mapping method. In this method, the read is positioned on the reference genome based on known exon annotations. Thus, this method cannot be used to identify a new splice site. QPALMA adopts the machine learning strategy and trains support vector machines for site identification using known splice sites. Vmatch has been adopted for positioning. However, because the efficiency of Vmatch is not high enough compared with Bowtie, Vmatch is not used for comparing reads. TopHat first positions the sequence on the reference genome using Bowtie. MAQ successfully positions the sequence assembly on the reference genome. Then, a possible splice site is recognized based on the adjacent exons. Additionally, the sequences not positioned on the reference genome are collected to establish the vacancy seed index. Finally, the vacancy expansion is compared in order to obtain the possible splice sites. According to a test reported by the authors, TopHat processed 2.2 million reads per hour, whereas QPALMA processed approximately 180,000. However, the performance will be poor when the depth of sequencing is low or the intron is very short because the algorithm adopts exon islands.

SpliceMap consists of four main steps: half-read mapping, seeding selection, site search, and paired-end filtering. First, SpliceMap splits the read into halves. Alignment positioning is performed between each portion and the gene sequence. Then, the remaining half is positioned on the downstream region within the range of the longest intron. This approach requires the read length to be at least 50 bp. Therefore, SpliceMap cannot process read lengths <50 bp.

When we compared SpliceMap with ERANGE, ERANGE discovered 160,899 sites, whereas SpliceMap accurately predicted 127,043 sites. Moreover, 24,274 of the 151,317 sites discovered by SpliceMap were not discovered by ERANGE, of which 23,020 represent new splice sites. However, these new sites are unconfirmed. The MapSplice software appeared after TopHat and SpliceMap. MapSplice is not based on the characteristics of splice sites or the length of an intron. It also has the potential to discover new sites and can adapt the length of the read.

The emergence of SOAPsplice improved the evaluation standard of splice site prediction software. SOAPsplice not only depends on the number of recognition splice sites but also emphasizes a high accuracy and low error rate. The experiment described in the next section revealed that the performance of SOAPsplice was comparatively outstanding. SplitSeek is strict with regard to the format of the input data and only supports data generated by ABI SOLiD. Moreover, because the input data are processed by a complete ABI transcriptome analysis tool, the application is not very wide. HMMSplicer is similar to SpliceMap but possesses several innovations. First, it divides the read into halves and compares halves with the genome sequence. The exon boundary (i.e., the $5'$ terminal) is obtained using the hidden Markov model (HMM). Second, the remaining half is positioned downstream the first half to determine the boundary $3'$ terminal of the intron. Both common (GT-AG, GC-AG, and AT-AC) and uncommon splice sites are recorded during this process. Finally, the scores of candidate loci are graded using the scoring algorithm.

3.3. Aligning Spliced Reads to the Reference Genome. Read lengths generated by all types of sequencing platforms are growing concomitant with the development of deep sequencing and RNA-seq technology. In the early days, read lengths were usually approximately 32 bp, and most of the software programs did not consider the location of the spliced reads on the reference genome. However, with the generation of longer reads, new requirements were put forward for locating software.

Reads mapping and alternative splicing detection are two steps in an analysis workflow. RNA read alignment is the precursor step and splice isoform detection is the successor step. Splice isoform detection tools include Cufflinks [81] and Scripture [82]. Cufflinks is a software tool for detecting the specific expression genes. If users have two groups of RNA-Seq data, such as ill and normal persons, it would be better to employ Cufflinks for the key genes detection. Scripture is a method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome *ab initio*.

Researchers applied the preprepared splice site database when they began trying to align spliced reads to the reference genome. However, the existing annotation of the transcriptome was far from being perfect. Therefore, some researchers once again began using BLAT to locate reads.

The TopHat software program solved these problems and thus became widely used by researchers; moreover, its vision has been expanding in every release from its initial release. In addition to its ability to align spliced reads to the reference genome, TopHat can also predict possible splice sites. These splice sites play an important role in improving the annotation of the transcriptome. The initial vision of TopHat had many limitations; however, the adoption of new methods in the software updates has improved TopHat's performance.

With the development of sequencing technologies, reads with lengths >100 bp have been produced on a large scale. These reads may span one or more spliced sites, which introduces difficulty in aligning spliced reads. The SpliceMap software is capable of processing longer reads (read lengths > 50 bp). To process these long reads, SpliceMap divides the reads into overlapping short read fragments. Then, they are annotated with the locating information of whole reads based on the locating information of the short read fragments.

MapSplice is another package that aligns spliced reads to the reference genome, although it applies a different method. The MapSplice algorithm is suitable for all types of read lengths. It is similar to SpliceMap in that it does not use continuous aligning of the reads to create an exon library in advance. Because the MapSplice package does not depend on spliced read signal information when aligning reads, it can locate some reads that SpliceMap cannot align. It can also be used to predict new spliced reads with no spliced read signal information. Another advantage of the MapSplice package is its high efficiency compared with most other software.

Package SeqSaw was proposed by Wang et al. [83] and is totally different from TopHat and MapSplice. It was similar to the SpliceMap package in its early releases. However, SeqSaw use has dynamically changed to Hash Table to reduce the search space. The core algorithm of SeqSaw is focused on locating short reads to the genome. There are very few introns >400 Kb in the known mammalian genome. Thus, we can define intron lengths as being less than a certain value, with a default value of 400 Kb. Users can adjust the value according to the needs of different species or datasets. However, SeqSaw uses certain means and performs a large amount of optimization, which greatly reduces the search space.

The R package DEGseq [83] has been proposed to detect small changes in the genetic expression of each sample. It is used to assess the trend of background noise in MA due to technological repeats. Figure 2 shows the working process of DEGseq.

The difference between a DNA aligner and an RNA aligner is that an RNA aligner can tolerate extra-long deletions (introns) while DNA aligners cannot [84]. Moreover, many RNA aligners are constructed based on DNA aligners (i.e., TopHat is built based on Bowtie). STAR is the latest and most popular RNA-seq alignment tools. In addition to unbiased de novo detection of canonical junctions, STAR can discover noncanonical splices and chimeric (fusion) transcripts and is also capable of mapping full-length RNA sequences [85].

4. Experiments Using State-of-the-Art Software Tools

HMMSplicer, SOAPsplice, TopHat, and STAR were used to perform the following analysis of Illumina/Solexa output

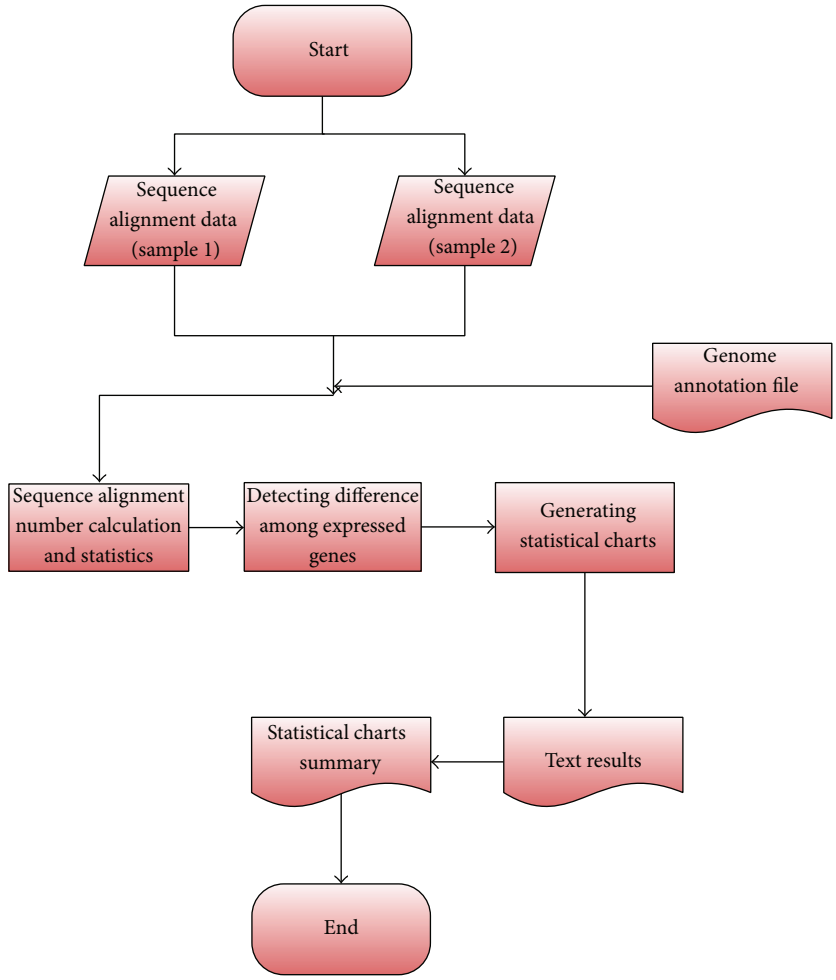

Figure 2: Working process of DEGseq.

data. The reference genome data are from the tenth human chromosome. The gene sequence was processed into RNA-seq sequences with different read lengths and different sequencing depths as the test data for SOAPsplice and TopHat. HMMSplicer does not support double-end sequencing data, so each pair of FASTQ data was merged into a FASTQ file as the test data for HMMSplicer.

Figure 3 shows that, in the premise of the 50 bp read length, each type of software predicts an increase in the number of loci that increases with the development of sequencing technologies. The accuracy of TopHat is poorer compared with the other two programs within a sequencing depth range of 1x to 10x, and the error rate is still high. The accuracy of TopHat increased rapidly after the sequencing was deepened. SOAPsplice and TopHat performed well in the aspect of accuracy, although the error rate was significantly worse for TopHat. STAR works best among the four tested tools. SOAPsplice and STAR performed well in both aspects.

5. Conclusion

In this study, we analyzed and compared the current AS-associated algorithms and software. We summarized the present situation of AS. The read mapping, including AS and site recognition algorithms, remained the focus of the current research. We aimed to improve the algorithm's quality in order to increase the number of prediction sites as much as possible and to meet the high-accuracy rate. RNA-seq data size is very large due to the continuous development of next-generation sequencing technology. This study represents a broad platform for AS and other fields of bioinformatics. This review of experimental and research methods for AS may be helpful for other researchers.

Although high-throughput sequencing has given rise to an unprecedented opportunity for the study of AS, few scholars study AS based on RNA-seq data. Therefore, the available algorithms and software are not rich compared with those based on EST/cDNA theory. Significant differences are found in the alignment step between the algorithms and the software using next-generation technology. This step represents the critical step based on the study of RNA-seq data. The software tools and algorithms need to be considered in parallel as the read data becomes more massive [86]. Genome-wide analysis will be the hot topic for all alternative and epigenetic research fields [87]. Moreover, many of the

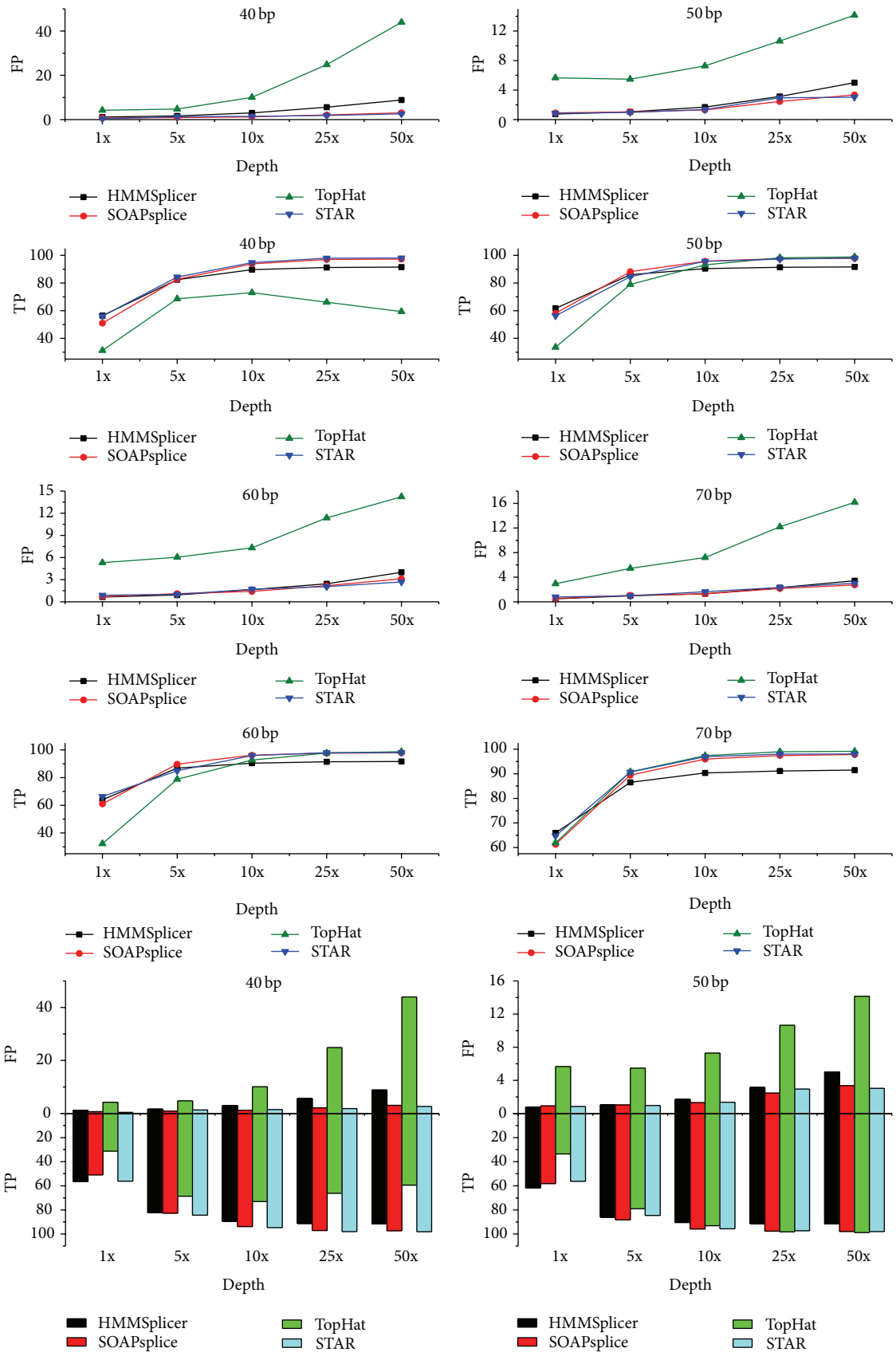

Figure 3: Continued.

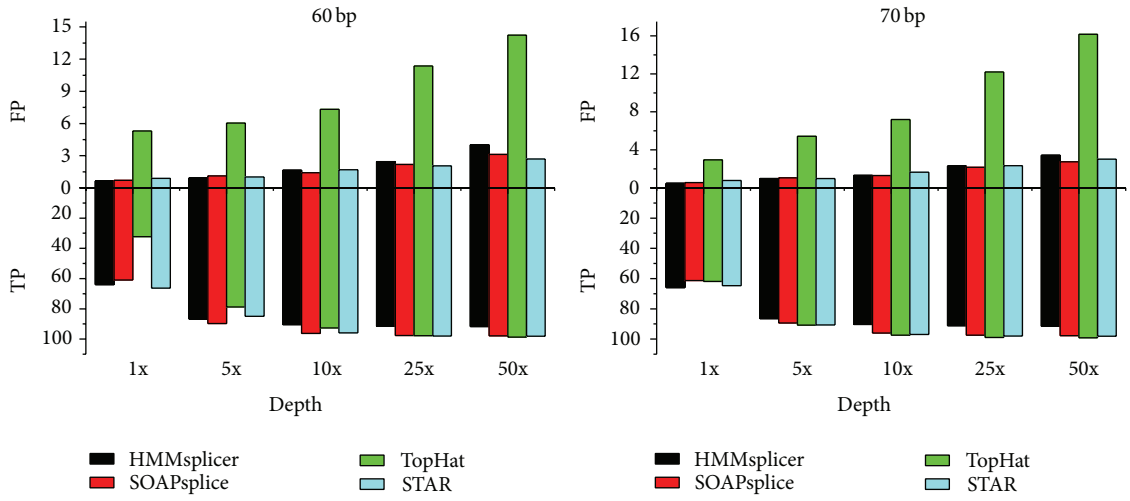

Figure 3: Comparison of HMMSplicer, SOAPsplice, STAR, and TopHat.

special databases based on RNA-seq data are not perfect. The corresponding new research methods and databases will be perfected with the constantly developing study of AS.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. M. Johnson, J. Castle, P. Garrett-Engele et al., "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science*, vol. 302, no. 5653, pp. 2141–2144, 2003.

[2] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.

[3] B. Modrek and C. Lee, "A genomic view of alternative splicing," *Nature Genetics*, vol. 30, no. 1, pp. 13–19, 2002.

[4] M. Dutertre, S. Vagner, and D. Auboeuf, "Alternative splicing and breast cancer," *RNA Biology*, vol. 7, no. 4, pp. 403–411, 2010.

[5] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[6] S. Hua, W. Yun, Z. Zhiqiang, and Q. Zou, "A discussion of microRNAs in cancers," *Current Bioinformatics*, vol. 9, no. 5, pp. 453–462, 2014.

[7] S. Stamm, J.-J. Riethoven, V. Le Texier et al., "ASD: a bioinformatics resource on alternative splicing," *Nucleic Acids Research*, vol. 34, pp. D46–D55, 2006.

[8] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[9] B. Liu, J. Yi, A. Sv et al., "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions," *BMC Genomics*, vol. 14, no. 8, article S3, 2013.

[10] L. T. Chow, R. E. Gelinas, T. R. Broker et al., "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA (Reprinted from Cell, vol 12, pg 1–12, 1977)," *Reviews in Medical Virology*, vol. 10, no. 6, pp. 362–369, 2000.

[11] M. T. Dimon, K. Sorber, and J. L. DeRisi, "HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data," *PLoS ONE*, vol. 5, no. 11, Article ID e13875, 2010.

[12] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, "Binmempredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[13] W. G. Fairbrother, R.-F. Yeh, P. A. Sharp, and C. B. Burge, "Predictive identification of exonic splicing enhancers in human genes," *Science*, vol. 297, no. 5583, pp. 1007–1013, 2002.

[14] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annual Review of Biochemistry*, vol. 72, pp. 291–336, 2003.

[15] P. Bonizzoni, R. Rizzi, and G. Pesole, "Computational methods for alternative splicing prediction," *Briefings in Functional Genomics and Proteomics*, vol. 5, no. 1, pp. 46–51, 2006.

[16] B. Modrek, A. Resch, C. Grasso, and C. Lee, "Genome-wide detection of alternative splicing in expressed sequences of human genes," *Nucleic Acids Research*, vol. 29, no. 13, pp. 2850–2859, 2001.

[17] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[18] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[19] L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller, "A computer program for aligning a cDNA sequence with a genomic DNA sequence," *Genome Research*, vol. 8, no. 9, pp. 967–974, 1998.

[20] N. Kim, S. Shin, and S. Lee, "ECgene: genome-based EST clustering and gene modeling for alternative splicing," *Genome Research*, vol. 15, no. 4, pp. 566–576, 2005.

[21] T. Castrignanò, R. Rizzi, I. G. Talamo et al., "ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization," *Nucleic Acids Research*, vol. 34, pp. W440–W443, 2006.

[22] S. J. Wheelan, D. M. Church, and J. M. Ostell, "Spidey: a tool for mRNA-to-genomic alignments," *Genome Research*, vol. 11, no. 11, pp. 1952–1957, 2001.

[23] J. Usuka, W. Zhu, and V. Brendel, "Optimal spliced alignment of homologous cDNA to a genomic DNA template," *Bioinformatics*, vol. 16, no. 3, pp. 203–211, 2000.

[24] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics*, vol. 21, no. 9, pp. 1859–1875, 2005.

[25] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, and W. Li, "A statistical method for the detection of alternative splicing using RNA-seq," *PLoS ONE*, vol. 5, no. 1, Article ID e8529, 2010.

[26] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[27] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[28] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, article S3, 2014.

[29] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[30] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[31] Q. Zou, Y. Mao, L. Hu, Y. Wu, and Z. Ji, "miRClassify: an advanced web server for miRNA family classification and annotation," *Computers in Biology and Medicine*, vol. 45, no. 1, pp. 157–160, 2014.

[32] B. Liu, X. Wang, L. Lin, B. Tang, and Q. Dong, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[33] Y. Zhang, B. Liu, Q. Dong, and V. X. Jin, "An improved profile-level domain linker propensity index for protein domain boundary prediction," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 7–16, 2011.

[34] G. Wang, K. Qi, Y. Zhao et al., "Identification of regulatory regions of bidirectional genes in cervical cancer," *BMC Medical Genomics*, vol. 6, no. 1, article S5, 2013.

[35] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, no. 1, article 298, 2014.

[36] B. Liu, J. Xu, X. Lan et al., "iDNA-Prot—dis: identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Article ID e106691, 2014.

[37] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[38] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[39] X.-Y. Cheng, W.-J. Huang, S.-C. Hu et al., "A global characterization and identification of multifunctional enzymes," *PLoS ONE*, vol. 7, no. 6, Article ID e38979, 2012.

[40] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[41] L. Cheng, Z.-G. Hou, Y. Lin, M. Tan, W. C. Zhang, and F.-X. Wu, "Recurrent neural network for non-smooth convex optimization problems with application to the identification of genetic regulatory networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 5, pp. 714–726, 2011.

[42] P. M. Feng, H. Lin, W. Chen, and Y. Zuo, "Predicting the types of j-proteins using clustered amino acids," *BioMed Research International*, vol. 2014, Article ID 935719, 8 pages, 2014.

[43] P. M. Feng, W. Chen, H. Lin, and K.-C. Chou, "IHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[44] W. Chen, P. Feng, and H. Lin, "Prediction of replication origins by calculating DNA structural properties," *FEBS Letters*, vol. 586, no. 6, pp. 934–938, 2012.

[45] W. C. Li, J. Z. Zhong, P. P. Zhu et al., "Sequence analysis of origins of replication in the *Saccharomyces cerevisiae* genomes," *Frontiers in Microbiology*, vol. 5, article 574, 2014.

[46] W. Chen, H. Lin, and P. M. Feng, "DNA physical parameters modulate nucleosome positioning in the *Saccharomyces cerevisiae* genome," *Current Bioinformatics*, vol. 9, no. 2, pp. 188–193, 2014.

[47] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo, and K.-C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[48] S.-H. Guo, E.-Z. Deng, L.-Q. Xu et al., "INuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[49] P. M. Feng, W. Chen, and H. Lin, "Prediction of CpG island methylation status by intergrating DNA physicochemical properties," *Genomics*, vol. 104, no. 4, pp. 229–233, 2014.

[50] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, and K.-C. Chou, "iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition," *Analytical Biochemistry*, vol. 462, pp. 76–83, 2014.

[51] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[52] Z. Yu, H. Chen, J. You et al., "Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 727–740, 2014.

[53] Z. Yu, L. Li, J. You, H.-S. Wong, and G. Han, "SC^3: triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1751–1765, 2012.

[54] W. Chen, H. Lin, P. Feng, and J. Wang, "Exon skipping event prediction based on histone modifications," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 241–249, 2014.

[55] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition," *BioMed Research International*, vol. 2014, Article ID 623149, 12 pages, 2014.

[56] Y. Q. Xing, L. R. Zhang, and L. F. Luo, "Prediction of alternative splicing sites of cassette exons and intron retention in human genome," *Acta Biophysica Sinica*, vol. 24, pp. 393–401, 2008.

[57] M. Wang and A. Marín, "Characterization and prediction of alternative splice sites," *Gene*, vol. 366, no. 2, pp. 219–227, 2006.

[58] A. Krogh, "Using database matches with HMMGene for automated gene detection in *Drosophila*," *Genome Research*, vol. 10, no. 4, pp. 523–528, 2000.

[59] S. Brunak, J. Engelbrecht, and S. Knudsen, "Prediction of human mRNA donor and acceptor sites from the DNA sequence," *Journal of Molecular Biology*, vol. 220, no. 1, pp. 49–65, 1991.

[60] S. M. Hebsgaard, P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouzé, and S. Brunak, "Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information," *Nucleic Acids Research*, vol. 24, no. 17, pp. 3439–3452, 1996.

[61] G. Parra, E. Blanco, and R. Guigó, "GeneID in *Drosophila*," *Genome Research*, vol. 10, no. 4, pp. 511–515, 2000.

[62] M. Pertea, X. Lin, and S. L. Salzberg, "GeneSplicer: a new computational method for splice site prediction," *Nucleic Acids Research*, vol. 29, no. 5, pp. 1185–1190, 2001.

[63] S. Degroeve, Y. Saeys, B. de Baets, P. Rouzé, and Y. van de Peer, "SpliceMachine: Predicting splice sites from high-dimensional local context representations," *Bioinformatics*, vol. 21, no. 8, pp. 1332–1338, 2005.

[64] Y. Xiaoling and S. S. Tang Tian, "Research progress and application of next-generation sequencing," *Biotechnology Bulletin*, vol. 10, pp. 76–80, 2010.

[65] K. F. Au, H. Jiang, L. Lin, Y. Xing, and W. H. Wong, "Detection of splice junctions from paired-end RNA-seq data by SpliceMap," *Nucleic Acids Research*, vol. 38, no. 14, pp. 4570–4578, 2010.

[66] X. Wang, X.-W. Wang, L.-K. Wang, Z.-X. Feng, and X.-G. Zhang, "A review on the processing and analysis of next-generation RNA-seq data," *Progress in Biochemistry and Biophysics*, vol. 37, no. 8, pp. 834–846, 2010.

[67] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[68] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: an alignment tool for large scale genome resequencing," *PLoS ONE*, vol. 4, no. 11, Article ID e7767, 2009.

[69] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno, "SHRiMP: accurate mapping of short color-space reads," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000386, 2009.

[70] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[71] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[72] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[73] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[74] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Research*, vol. 38, no. 6, pp. 1767–1771, 2009.

[75] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing," *Genome Research*, vol. 17, no. 11, pp. 1697–1706, 2007.

[76] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, 2007.

[77] J. Butler, I. MacCallum, M. Kleber et al., "ALLPATHS: de novo assembly of whole-genome shotgun microreads," *Genome Research*, vol. 18, no. 5, pp. 810–820, 2008.

[78] F. de Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *Bioinformatics*, vol. 24, no. 16, pp. i174–i180, 2008.

[79] K. Wang, D. Singh, Z. Zeng et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic acids research*, vol. 38, no. 18, article e178, 2010.

[80] A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten, "Global and unbiased detection of splice junctions from RNA-seq data," *Genome Biology*, vol. 11, no. 3, article r34, 2010.

[81] C. Trapnell, A. Roberts, L. Goff et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, 2012.

[82] M. Guttman, M. Garber, J. Z. Levin et al., "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, no. 5, pp. 503–510, 2010.

[83] L. Wang, X. Wang, X. Wang, Y. Liang, and X. Zhang, "Observations on novel splice junctions from RNA sequencing data," *Biochemical and Biophysical Research Communications*, vol. 409, no. 2, pp. 299–303, 2011.

[84] T. Steijger, J. F. Abril, P. G. Engström et al., "Assessment of transcript reconstruction methods for RNA-seq," *Nature Methods*, vol. 10, no. 12, pp. 1177–1184, 2013.

[85] A. Dobin, C. A. Davis, F. Schlesinger et al., "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[86] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 637–647, 2014.

[87] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings in Functional Genomics*, 2014.