Genomics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer

Qi Liu [a], Chong Chen [a], Enjian Shen [a], Fangqing Zhao [b], Zhongsheng Sun [a,b,*], Jinyu Wu [a,**]

[a] Institute of Genomic Medicine, Wenzhou Medical College, Wenzhou 325035, China
[b] Beijing Institutes of Life Science, Chinese Academy of Science, Beijing 100101, China

ARTICLE INFO

Article history:
Received 5 August 2011
Accepted 16 December 2011
Available online xxxx

Keywords:
Next-generation sequencing
Transcriptome
RNA-Seq
Alternative splicing
Soft
Visualization

ABSTRACT

Alternative splicing is a crucial mechanism by which diverse gene products can be generated from a limited number of genes, and is thought to be involved in complex orchestration of eukaryotic gene expression. Next-generation sequencing technologies, with reduced time and cost, provide unprecedented opportunities for deep interrogation of alternative splicing at the genome-wide scale. In this study, an integrated software SplicingViewer has been developed for unambiguous detection, annotation and visualization of splice junctions and alternative splicing events from RNA-Seq data. Specifically, it allows easy identification and characterization of splice junctions, and holds a versatile computational pipeline for in-depth annotation and classification of alternative splicing with different patterns. Moreover, it provides a user-friendly environment in which an alternative splicing landscape can be displayed in a straightforward and flexible manner. In conclusion, SplicingViewer can be widely used for studying alternative splicing easily and efficiently. SplicingViewer can be freely accessed at http://bioinformatics.zj.cn/splicingviewer.

1. Introduction

Alternative splicing is a highly regulated and conserved process by which the exons of pre-mRNA are reconnected in a number of ways to form different mRNAs that are then translated into protein isoforms. Therefore, a limited number of genes in an organism's genome are able to form a large complex proteome [1]. Alternative splicing is a notable phenomenon commonly found in eukaryotes that is empirically classified into several different patterns according to the structures of exons, including skipped exons, mutually exclusive exons, alternative 5′ splice sites, alternative 3′ splice sites, alternative first exons, alternative last exons and retained introns [1,2]. Aberrant splicing has been reported to be linked to many diseases, such as spinal muscular atrophy, tauopathies, and Hutchinson–Gilford progeria syndrome. [3]. Thus, the detection and annotation of alternative splicing are essential to better understand cellular process and development in biological and medical research. Recently, RNA-Seq, a powerful and rapidly evolving high-throughput sequencing technology, has opened new horizons in understanding the transcriptome [4–6].

To study alternative splicing from RNA-Seq data, a number of useful tools have been introduced. Tophat is an efficient pipeline that can identify splice junctions in large-scale mapping of RNA-Seq reads [7]. QPalma is a junction alignment tool that maximizes alignment accuracy by using optimum parameters that are estimated based on SVM-like algorithm [8]. SpliceMap is an implementation of split-read alignment algorithm using the mapping of half-reads to identify locations of junctions [9]. Supersplat is a method for unbiased splice junction detection in empirical RNA-Seq data [10]. MMES [11], a new statistical metric, is developed to not only measure the quality of junction reads but also implement empirical statistical models for detection of exon junctions. MapNext is a software tool for spliced and unspliced alignments and detection in short read sequences [12]. In addition, other splicing detection tools were newly developed recently, including SplitSeek [13], MapSplice [14], HMMSplicer [15], NSMAP [16] and SOAPsplice (http://soap.genomics.org.cn/soapsplice.html).

In this study, an integrated tool SplicingViewer has been developed to enable users to detect splice junctions, annotate alternative splicing events, and visualize alternative splicing patterns. Additionally, SplicingViewer can also efficiently display genome mapping reads and junction mapping reads at high speed and low memory cost. In conclusion, SplicingViewer is, to our knowledge, the first software dedicated not only to detection of splice junctions, but also to annotation, visualization and manipulation of alternative splicing events.

2. Results

2.1. Validation and parameter evaluation

To evaluate the performance of SplicingViewer, we simulated Illumina sequencing short reads from 92,543 transcripts of the human

* Correspondence to: Z. Sun, Institute of Genomic Medicine, Wenzhou Medical College, Wenzhou 325035, China. Fax: +86 577 88831309.
** Corresponding author. Fax: +86 577 88831309.
E-mail addresses: sunzs@psych.ac.cn (Z. Sun), iamwujy@yahoo.com.cn (J. Wu).

0888-7543/$ – see front matter © 2011 Published by Elsevier Inc.
doi:10.1016/j.ygeno.2011.12.003

genome in ASTD (the Alternative Splicing and Transcript Diversity database) [17] at various sequencing depths. Initially, the performance of different short read aligners in splice junction calling was evaluated (Fig. 2A). For all the aligners evaluated, including MAQ [18], BWA [19], Bowtie [20], and SOAP2 [21], the sensitivity shows a similar ascending trend when the short read depth increases from 1× to 20×, reaching the highest (>99%) value at a depth of 20×, and remains relatively constant when the depth is >20×. In contrast, the PPVs (positive predictive value) of all aligners exhibit a descending trend when the depth is >20× (Fig. 2B). This can be explained by the fact that increased false mappings will be generated when more reads are introduced. MAQ and BWA were noted to have a relatively higher performance among the four aligners. However, BWA can also support multiple threads and have a higher speed. Therefore, BWA was chosen as the aligner in the following analyses.

In order to evaluate the ability of SplicingViewer to identify junctions with different read lengths, we simulated short reads with different lengths at a depth of 20×. The sensitivity of SplicingViewer remains relatively constant when the read length is <75 bp, but shows a decreasing trend when the read length is >75 bp (Fig. 2C). However, the PPV increases with the read length (Fig. 2D). The minimum number of reads needed to support the junction is a critical parameter in SplicingViewer. We evaluated the performance of SplicingViewer using different values of this parameter (Figs. 2E and F). Sensitivity sharply decreases when the value is >4. However, the PPV reaches ~99% when the value is 2 and remains steady when the value goes up from 2 to 10. Accordingly, we set the default minimum junction read number parameter to 2.

2.2. Implementation and visualization

The splice junction detection and alternative splicing event annotation are integrated into one command line program containing two steps that is written in JAVA programming language. Based on the sorted short read mapping BAM file, reference genome sequence and the known gene models, the first step generates splice junctions, and the second step generates the annotated patterns of alternative splicing that can be used for the visualization in SplicingViewer.

The inputs for the visualization in SplicingViewer include: 1) the reference genome sequence, 2) sorted BAM file of genome mapping, 3) sorted BAM file of junction mapping, 4) the known gene models, and 5) the annotation result of alternative splicing patterns obtained from the command line program.

SplicingViewer offers a user-friendly interface and can be used in a platform-independent manner (Fig. 3). Patterns of alternative splicing are displayed with different shapes, with rectangle and polygonal lines representing the exon and junction, respectively. When the mouse hovers on and clicks a polygonal line, the short reads supporting that junction will be displayed in a popup dialog. All short reads, including reads mapped to the reference genome and the junctions, are optimally placed in multiple lines with a compact arrangement that can be visualized intuitively. In order to get a better view, SplicingViewer also provides users with many adjustable options, including the zoom in/out option, which can be used to adjust the alignment and annotation view resolution, and the font, color, and shape options, which can be used to change the appearance of the visualization. In addition, SplicingViewer has a query system, which allows users to search gene models and alternative splicing events by their IDs.

2.3. Computer performance

The computational performance of SplicingViewer was evaluated in two steps. Firstly, we tested the performance of command line program on RNA-Seq data from LNCaP prostate cancer (30,784,244 pair-end short reads, 75 bp). The command line program was run on a HP

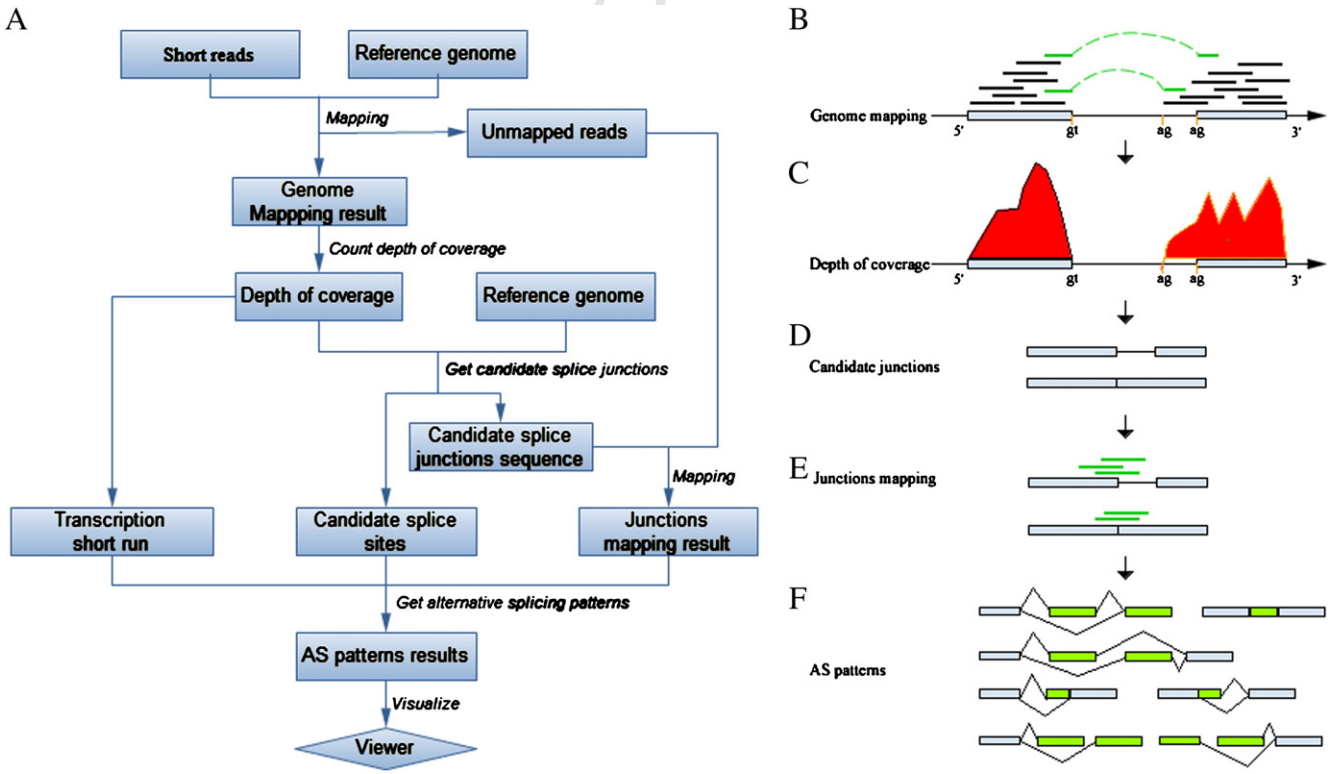

Fig. 1. The schematic diagram of SplicingViewer pipeline. (A) The overall procedure of SplicingViewer. (B) Mapping of short reads to the reference genome. Black short lines represent the reads mapped to the genome sequence, while unmapped short reads are represented by short lines colored green. (C) The coverage calculation of the loci on the genome inside the gene bounds. (D) Detection of the candidate splice sites and splice junctions. (E) Mapping of unmapped reads to the junction sequences to detect the real splice junctions. (F) Pattern annotation of alternative splicing events. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: Q. Liu, et al., Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer, Genomics (2012), doi:10.1016/j.ygeno.2011.12.003

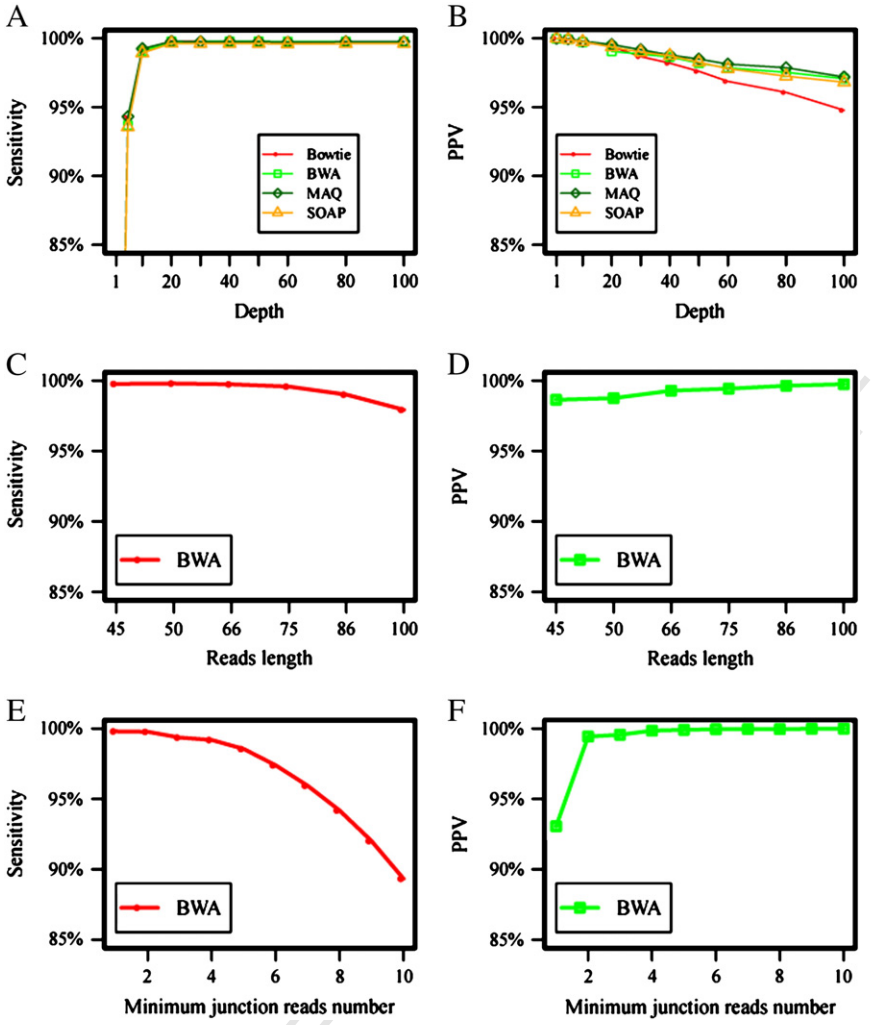

Fig. 2. Validation and parameter evaluation. (A) Sensitivity of different aligners in real splice junction detection at different short read depths (ranging from 1× to 100×). (B) PPVs of different aligners in real splice junction detection at different sequencing depths. (C) Sensitivity in real splice junction detection, measured by mapping simulated short reads with different sequence lengths. (D) PPV in real splice junction detection, measured by mapping simulated short reads with different sequence lengths. (E) Sensitivity in real splice junction detection with different thresholds of minimum junction read number. (F) PPV in real splice junction detection with different thresholds of minimum junction read number.

cluster work station (AMD 64-bit 2.2 GHz CPU, 32 GB of RAM) with the result of mapping short reads to reference genome as input. The mapping result is in BAM format and contains 20,558,742 mapped reads and 10,225,502 unmapped reads. It took the program approximately 5 h and a 5.5 GB memory to finish the analysis and obtain the final annotated patterns of alternative splicing. Then, the visualization process was tested on a typical Windows XP system with Intel Core 2 Duo E7400 (2.80 GHz) and 2 GB memory. It took SplicingViewer 4.156 s and 120 MB of RAM to load the data. After loading the data, SplicingViewer ran smoothly in displaying short reads, alternative splicing patterns and various other kinds of important information.

3. Discussion

Compared with traditional methods, such as microarray or aligning gene sequences against EST sequences, RNA-Seq shows high superiority because it can be used for comprehensive transcription profiling and deciphering the mechanisms of alternative splicing with unprecedented accuracy and at a low error rate. To provide an integrated framework for identification, in-depth annotations of alternative splicing events and visualization of alternative splicing patterns from RNA-Seq data, we have developed SplicingViewer. As an ongoing project, SplicingViewer will in the future allow users to perform the comparison of alternative splicing events among multiple samples. Additionally, SplicingViewer will be further improved according to users' feedbacks and suggestions. Taken together, SplicingViewer, a promising tool for globally characterizing the alternative splicing landscape of the transcriptome, would enhance our understanding of transcriptome complexity and accelerate research progress in biological and medical areas.

4. Materials and methods

In brief, SplicingViewer uses three main steps to deeply survey alternative splicing complexity from RNA-Seq data: 1) alignment of short reads to reference genome, 2) detection of candidate splice junctions, and 3) alignment of unmapped short reads to splice junctions (Fig. 1A).

Firstly, short reads are mapped to the reference genome using short read aligners, including MAQ [18], BWA [19], Bowtie [20] and SOAP2 [21] (Fig. 1B). The mapping results are then converted into SAM/BAM format with SAMtools [22], the common input format used in the following analyses. Using GATK (the Genome Analysis Toolkit) [23], depth of coverage at each locus of the known gene model is calculated based on the genome mapping records of short

Please cite this article as: Q. Liu, et al., Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer, Genomics (2012), doi:10.1016/j.ygeno.2011.12.003

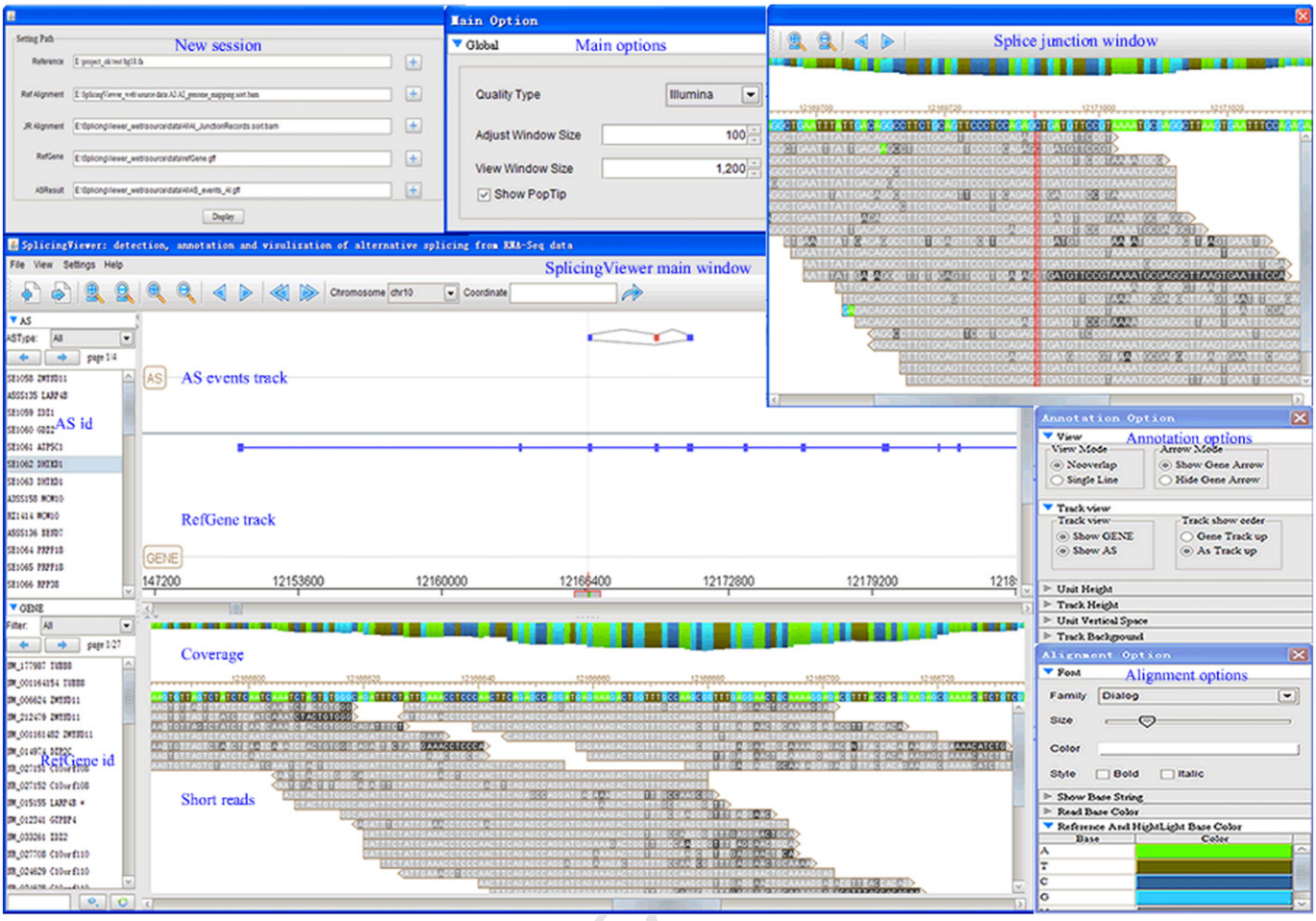

Fig. 3. Screenshots of SplicingViewer windows.

reads in BAM file (Fig. 1C). The short reads that are not mapped to the genome will be used in the following splice junction alignment.

Then, the reference genome sequence, together with the known gene models, is used to determine the candidate splice junctions according to the splice site rule. Among the canonical splice site pairs (donor–acceptor), GT-AG accounts for 98.3% with GC-AG and AT-AC accounting for only 1.5% and 0.2%, respectively [24]. In SplicingViewer, GC-AG and AT-AC can also be detected according to users' requirement. Let $I=$, be the collection of introns in gene model G, where n is the number of introns. The splice donor and acceptor sites in splice site pair will be searched separately on each element of I. For intron I_i, donor dinucleotide is searched along its sequence from 5′ end. Let s and d be the leading 5′ end dinucleotide site of I_i and the hit donor site, respectively. The genomic coordinates of s and d will be represented by S_s and S_d, respectively. Let D_i represent the genomic coordinate i on the genome sequence, then the coverage rate of the genomic intron region from S_s to S_d will be

$$M_d = \frac{\sum_{S_s \le i \le S_d} D_i}{S_d - S_s + 1}.$$

For the hit donor site d, if S_d is not equal to S_s and $M_d \ge 0.98$, indicating that the genomic intron region between s and d is covered by short reads with a small tolerance, then the search will move forward; and if S_d is not equal to S_s and $M_d < 0.98$, indicating that the hit site d resides at the border of the intron region covered by the short reads, then, to avoid missing candidate splice sites, the search will be extended forward with a distance equal to the length of short read and then terminated. However, when S_d is not equal to S_s and $M_d < 0.98$, if d is adjacent to s, indicating that there is no short read covering the intron region between s and d then the search will be terminated immediately.

Let $d_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,n}\}$ be all hit donor sites of I_i, where n represents the number of hit donor sites, we then add s to d_i, forming $d_i' = \{s, d_{i,1}, d_{i,2}, \ldots, d_{i,n}\}$. Let d_i^r be the potential donor sites of I_i. If $|d_i'| = 1$, we set $d_i^r = d_i'$; otherwise $d_i^r = \{s, d_{i,t-2}, d_{i,t-1}, d_{i,t}, \ldots, d_{i,n}\}$, where $d_{i,t}$ is the hit donor site which resides at the border of the 5′ intron region covered by the short reads.

The Acceptor dinucleotide is searched on the sequence of I_i from the 3′ end. Let e and a be the leading 3′ end dinucleotide site of I_i and the hit acceptor site, respectively. The genomic coordinates of e and a will be represented by S_e and S_a, respectively. The coverage rate of the genomic region from S_e to S_a will be

$$M_a = \frac{\sum_{S_a \le i \le S_e} D_i}{S_e - S_a + 1}.$$

The search criteria are set to be the same as those of the donor site. $a_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,n}\}$, which represents that all the hit acceptor sites of I_i, will be generated. Let $a_i' = \{e, a_{i,1}, a_{i,2}, \ldots, a_{i,n}\}$, then the potential acceptor site set of I_i, represented by a_i^r is set to be equal to a_i' if $|a_i'| = 1$; otherwise, $a_i^r = \{e, a_{i,t-2}, a_{i,t-1}, a_{i,t}, \ldots, a_{i,n}\}$, where $a_{i,t}$ is the hit acceptor site which resides at the border of the 3′ intron region covered by the short reads.

Let $J_i = \{d_i^r, a_i^r\}$ be the splice site set of I_i. $J = \{J_1, J_2, \ldots, J_n\}$ represents all splice sites of the intron set I in gene model G, and all candidate splice junctions of gene model G will be generated by grouping all donor–acceptor site pairs (GT-AG/GC-AG/AT-AC) from d_i^r in J_i and a_i^r

Please cite this article as: Q. Liu, et al., Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer, Genomics (2012), doi:10.1016/j.ygeno.2011.12.003

in J_j with all combinations of J_i and J_j, where $1 \le i \le n-1$ and $i<j \le$ n. In each donor–acceptor site pair, genome sequences upstream of the donor site coordinate and downstream of the acceptor site also coordinate, both of which are equal in length to the read, are cut off and joined to form a continuous splice junction sequence (Fig. 1D). This process is applied to all gene models, and the candidate splice junction sequences of each gene model will be generated.

Finally, the unmapped short reads in the genome mapping step are mapped to the splice junction sequences to identify the real splice junctions (Fig. 1E). By default, a real junction site must be supported by at least two unambiguously mapped short reads with a non-repetitive match position. All the splice junctions identified, together with the splice junction site information, are used to annotate the patterns of alternative splicing (Fig. 1F).

Acknowledgments

This work is supported by the National Natural Science Foundation of China (31171236/C060503).

References

[1] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature 456 (2008) 470–476.
[2] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, B.J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet. 40 (2008) 1413–1415.
[3] J. Tazi, N. Bakkour, S. Stamm, Alternative splicing and disease, Biochim. Biophys. Acta 1792 (2009) 14–26.
[4] S. Marguerat, J. Bahler, RNA-Seq: from technology to biology. Cell. Mol. Life Sci. 67, 569–579.
[5] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[6] M.F. Berger, J.Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L.A. Johnson, J. Robinson, R.G. Verhaak, C. Sougnez, R.C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D.E. Fisher, G. Getz, M. Meyerson, D.B. Jaffe, S.B. Gabriel, E.S. Lander, R. Dummer, A. Gnirke, C. Nusbaum, L.A. Garraway, Integrative analysis of the melanoma transcriptome. Genome Res. 20, 413–427.
[7] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.
[8] F. De Bona, S. Ossowski, K. Schneeberger, G. Ratsch, Optimal spliced alignments of short sequence reads, Bioinformatics 24 (2008) i174–i180.
[9] K.F. Au, H. Jiang, L. Lin, Y. Xing, W.H. Wong, Detection of splice junctions from paired-end RNA-Seq data by SpliceMap. Nucleic Acids Res. 38, 4570–4578.
[10] D.W. Bryant, Jr., R. Shen, H.D. Priest, W.K. Wong, T.C. Mockler, Supersplat-spliced RNA-Seq alignment. Bioinformatics 26, 1500–1505.
[11] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, W. Li, A statistical method for the detection of alternative splicing using RNA-Seq. PLoS One 5, e8529.
[12] H. Bao, Y. Xiong, H. Guo, R. Zhou, X. Lu, Z. Yang, Y. Zhong, S. Shi, MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads, BMC Genomics 10 (Suppl. 3) (2009) S13.
[13] A. Ameur, A. Wetterbom, L. Feuk, U. Gyllensten, Global and unbiased detection of splice junctions from RNA-Seq data. Genome Biol. 11, R34.
[14] K. Wang, D. Singh, Z. Zeng, S.J. Coleman, Y. Huang, G.L. Savich, X. He, P. Mieczkowski, S.A. Grimm, C.M. Perou, J.N. MacLeod, D.Y. Chiang, J.F. Prins, J. Liu, MapSplice: accurate mapping of RNA-Seq reads for splice junction discovery. Nucleic Acids Res. 38, e178.
[15] M.T. Dimon, K. Sorber, J.L. DeRisi, HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. PLoS One 5, e13875.
[16] Z. Xia, J. Wen, C.C. Chang, X. Zhou, NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq. BMC Bioinformatics 12, 162.
[17] G. Koscielny, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, J.J. Riethoven, F. Nardone, E. Stanley, C. Fallsehr, O. Hofmann, M. Kull, E. Harrington, S. Boue, E. Eyras, M. Plass, F. Lopez, W. Ritchie, V. Moucadel, T. Ara, H. Pospisil, A. Herrmann, G.R. J., R. Guigo, P. Bork, M.K. Doeberitz, J. Vilo, W. Hide, R. Apweiler, T.A. Thanaraj, D. Gautheret, ASTD: The Alternative Splicing and Transcript Diversity database, Genomics 93 (2009) 213–220.
[18] H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores, Genome Res. 18 (2008) 1851–1858.
[19] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[20] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[21] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics 25 (2009) 1966–1967.
[22] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[23] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.
[24] S. Stamm, J.J. Riethoven, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N.L. Barbosa-Morais, T.A. Thanaraj, ASD: a bioinformatics resource on alternative splicing, Nucleic Acids Res. 34 (2006) D46–D55.

Please cite this article as: Q. Liu, et al., Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer, Genomics (2012), doi:10.1016/j.ygeno.2011.12.003