BAsplice: Bi-direction Alignment for detecting splice junctions

Jingde Bu[1, 3†], Jiayan Wu[2†], Meili Chen[2], Jingfa Xiao[2], Jun Yu[2], Xuebin Chi[1], Zhong Jin[1*]
[1]Supercomputing Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
[3]Graduate University of Chinese Academy of Sciences, Beijing, China
[†] The first two authors contributed equally to this work.
[*]Correspondence to: Zhong Jin E-mail: zjin@sccas.cn

***Abstract*-- RNA-Seq is a revolutionary whole transcriptome shotgun sequencing technology performed by high-throughput sequencers, which provide more comprehensive information on differential expression of genes and benefit on novel splice variants identification. RNA-Seq reads is so short that it's a great challenge on mapping reads back to the reference effectively, especially when they span two or more exons. To improve the mapping efficiency, we introduce here a bi-direction alignment tool – BAsplice, which use RNA-Seq data to detect splice junctions without any additional information. Compare with another splice junction mapping software, SOAPsplice, BAsplice performs better in call rate and running time, but a little worse in accuracy.**
BAsplice is a free open-source software written in C language. It is available at https://github.com/vlcc/basplice.

Keywords-RNA-Seq; splice junctions; bi-direction alignment

I. INTRODUCTIONS

Recent development in high-throughput sequencing technology has caused the dramatically decreasing in sequencing costs, which makes the bioinformatics process to be the new bottleneck in the sequencing project. How to map the short reads back to the reference accurate and fast is a great challenge now. RNA-Seq [1] is a revolutionary whole transcriptome shotgun sequencing technology performed by high-throughput sequencers, which provide more comprehensive information on differential expression of genes and benefit on novel splice variants identification. As to alternative splicing events, RNA-Seq substitutes the previous mainstream methodology, expressed sequence tags (ESTs), with the low cost and high-throughput data. Splice junctions are points on a DNA sequence at which `superfluous' DNA is removed during the process of protein creation in higher organisms [2]. It is a basic and key task in an RNA-Seq experiment to find splice junctions by aligning reads to reference.

Most of the common aligners such as BWA [3] and Bowtie [4] do not align reads across splice junctions as efficient as they do in the exon regions, while spliced alignment is substantially necessary in genome structural annotation [5]. Several professional spliced alignment tools have been developed in recent including Q-PALMA [6], Tophat [7], MapSplie [8], SpliceMap [9] and SOAPsplice [10]. Q-PALMA applies machine-learning strategy to predict splice sites and uses known splice junctions from quality information and known alignments, which improves the accuracy but brings more bias by the training data. Tophat is an 'Exon-fist' methods [11] use a two-step process: firstly mapping reads to the reference by Bowtie and computing consensus by MAQ [12] from mapped reads; then generating splice junctions from neighboring exons and mapping initially unmapped reads (IUM reads) to the joint sites. On the premise of high sequencing coverage, Tophat is suitable for detecting splice junctions. SpliceMap splits IUM reads into two halves and maps both of them to the reference by SeqMap [13] or ELAND [14], then pick one as a seed to perform extension. It will filter alignment results by pair-ends information and canonical form of intron. MapSplice cuts IUM reads into several segments and maps them to the reference, then uses unmapped segments to do spliced alignment, finally merges all results and predicts splice junctions. The drawback of this tool is the mapping accuracy might decreases a little, when the read is so short and multiple aligned. SOAPsplice usually divides IUM reads into two segments to do the spliced alignment. When reads longer than 50 bp, it splits reads into multiple segments, and uses known splicing motifs to filter mapping result. Overviewing all these spliced alignment tools, the crucial point is how to segment IUM reads and how to find the spliced site from the segment alignment results.

In this investigation, we introduce a bi-direction alignment software special for detecting splice junctions, named BAsplice. It is a two-step approach which identifies splice junctions with both nonconservative and conservative motifs. First, BWA is used to align reads to the reference; second, BAsplice splits IUM reads into three segments equally and maps each of segments to the reference. Based on the mapping result and known splice junctions information, mapped segments will be extend to a propose position, which is thought to be junction site. Running time, call rate and accuracy are measured for performance,

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

compared with SOAPsplice. BAsplice shows better in the first two asepects.

II. RESULTS

A. Pipeline

The overview pipeline of BAsplice is shown in Figure 1. BAsplice takes a two-step approach to map RNA-Seq reads to the reference. Bi-direction BWT [15] is used to index the reference sequence, so we can search the sequence from both front and back ends. First, using BWA, a fast unspliced read aligner, to map reads to the reference genome, which we called Initial alignment; second, IUM reads generated from the first step process splice alignment. The details show as follow.

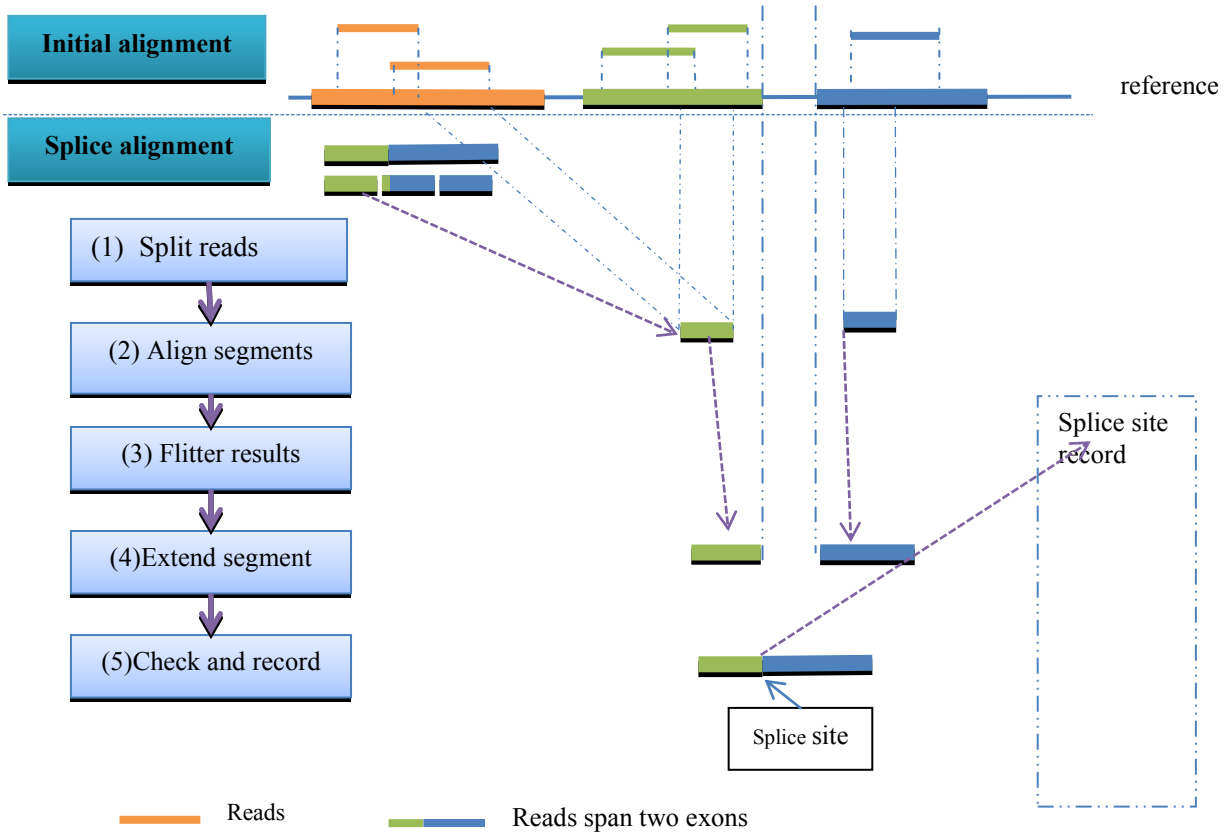

Fig. 1. BAsplice pipeline. This pipeline contains two phases: initial alignment and splice alignment. BWA is applied to align RNA-Seq reads to the reference, and IUM reads are collected in the first step. Then we divide each IUM reads into three segments, use BWA to align each segment, and filter the mapped results. One of filtered mapped segments is selected as a seed to process extension operation. Splice sites are verified based on features of splice junctions. All splice sites are recorded.

(1) Initial Alignment

In the first step, BAsplice aligns reads to the reference by BWA. BWA is a fast and accurate short reads aligner using FM-index [16] allows mismatches and gaps; the number of mismatch and gap can be configured by user themselves. Because of duplicate regions in gene, multiple alignments are allowed in BAsplice. The reads failed mapped in first step are treated as IUM reads, which are put into the second step, splice alignment.

(2) Splice Alignment

Splice alignment step in BAsplice contains five substeps (see Figure 1).

Substep 1: Split reads. All IUM reads generated from Initial alignment are divided into three segments equally. Due to the improvement of RNA-Seq methodology, three-cut segments are long enough to be unique mapped, and it can help increasing call rate.

Substep 2: Align segments. After split IUM reads, each segment is mapped to the reference by means of BWA. Mismatches are allowed here, but do not allow gaps. If only one segment in those three has mapped to the reference, we would treat this IUM read failed in splice alignment. If two of three have past, the read pass. This may derive from the very small exons that the read crosses more than two exons, or the read has too much error.

Substep 3: Filter results. The three-cut segment is so short that it may contain in many parts of the reference genome. Thus, we filter the segment alignment results with these two criteria:

a) Mapped segments should be on the same chromosome and on the same strand;

b) Distance of two segments which may be on two different exons should be in the scope of 50 to 50,000 bp [6].

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Substep 4: Extend segments. After the mapped segments (at least two) split from one IUM read have past the filter process, we select one of them as a seed and extend it to the right position.

Substep 5: Check and record. BAsplice designs a structure to record splice site information generating by searching the splice site near the retrieved position. It makes the split position of the read and the splice site confirm with each other. Another advantage of this record is helping us to judge the true splice site by hit number of all splice sites

B. Implementation

BAsplice is written in C language. It uses Bi-direction-BWT index the reference and uses BWA as initial and segments mapping software. Currently, it runs on Linux system.

C. Simulation results

We used chromosome 17 of human as the reference and extracted 5153 transcripts longer than 350 bp resulting in 17661 known splice junctions from Ensembl database. The simulated reads are generated using wgsim [17] with base error rate and mutation rate of 0.02 and 0.001. To test the read length effect, three different lengths of reads, 75, 100 and 150, are considered separately in the simulation. For each transcript, the reads are simulated at eleven kinds of coverage (0.1, 1, 5, 10, 20, 30, 40, 50, 60, 80 and 100 fold).

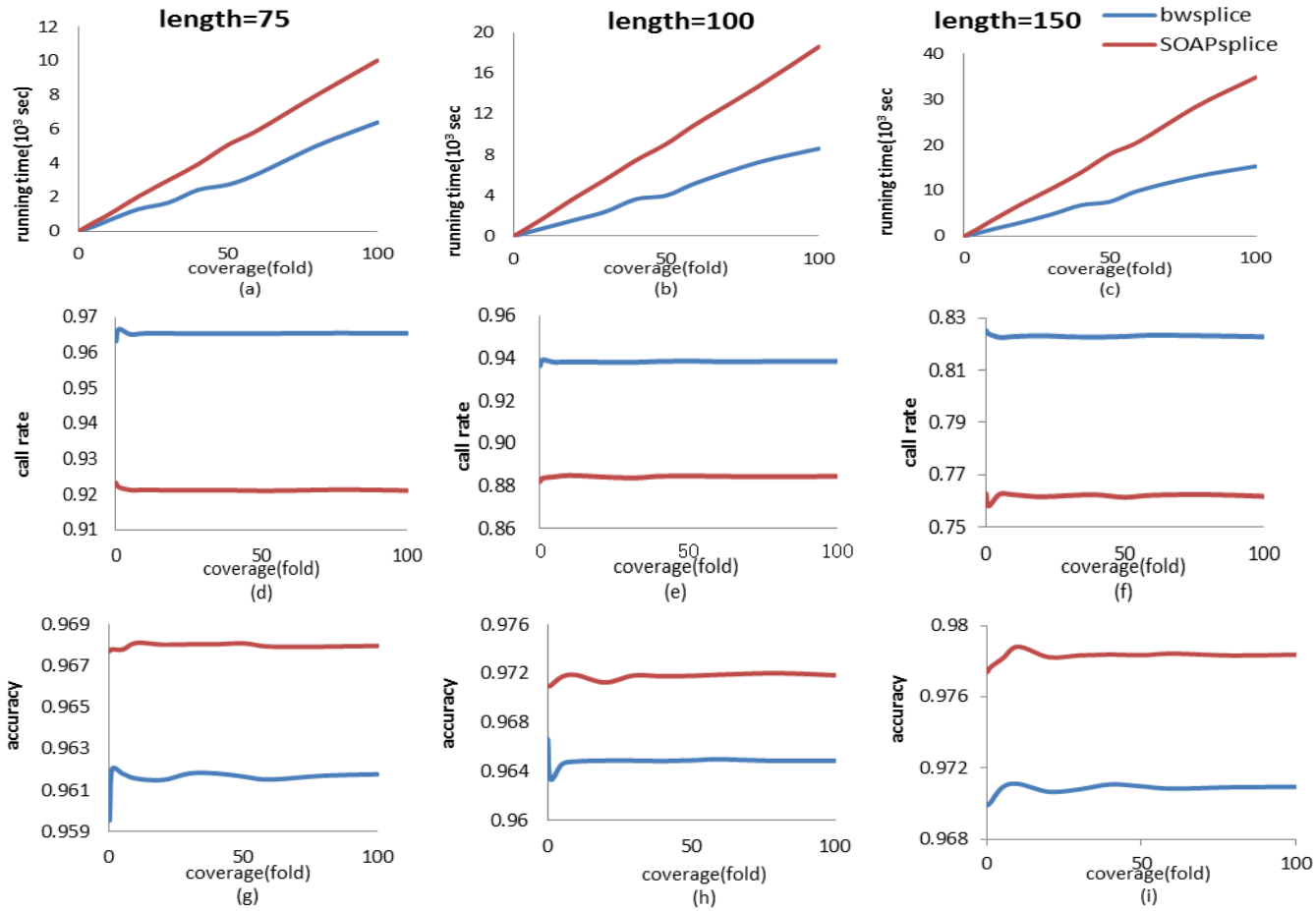

Fig. 2. The statistical results on 75, 100 and 150 bp simulated reads in BAsplice and SOAPsplice under different coverage of transcript. (a), (b) and (c) show the running time results. (d), (e) and (f) show the call rate results. (g), (h) and (i) show the accuracy result. Three columns are the statistical results of 75, 100 and 150 bp simulated reads.

Figure 2 shows the statistical results of running time, call rate and accuracy on 75, 100 and 150 bp simulated reads. Seen from the first line (a, b and c), the running time of BAsplice is about 2 times less than SOAPsplice in mapping simulated reads performed in the same platform (a single core of 2.8 G AMD Opteron 2220 processor, Centos 5.1 operation system). With longer sequence length, more time are saved by using BAsplice. The second line of this figure shows call rate results with different read length. BAsplice also performed better than SOAPsplice in this aspect. At the coverage of 50 fold, the call rate of SOAPsplice is about 92%, 88% and 76% on 75, 100 and 150 bp simulated reads, while the call rate of BAsplice on read lengths of 75, 100 and 150 is about 96.3%, 93.6% and 82.5%, separately. From Figure 2, it can be seen that, SOAPsplice's accuracy of called reads is 0.6 percent higher than BAsplice. When we check mapping errors, we find that the main reason is read mapped to reverse complement strand of the reference genome, in other words, the reference have duplicate regions on different strands. With the rise of read length

the call rate decrease, but the accuracy of mapped reads increase. These mainly because 25% of exons are shorter than 100 bp in human, read may span three or more exons when read length longer than 100 bp, this will bring about the failure of mapping to reference.

III. CONCLUSIONS

All detected splice sites information is record in splice site structure. All alignment are saved in files, another part of BAsplice can generate SAM format file [18] from it. This splice alignment software has been approved that it has a higher call rate but has a little lower accuracy of mapped read. The reason why time consumes decreased is that the number of iterations is reduced by shorter segments which are equally split from reads. Meanwhile, short segments may have more alignments, which led to the increase of call rate. Owing to the software only generates index of positive strand, the software will gives a false result if reads comes from negative strand but positive strand have the same region. When reads shorter than 70 bp, the segments are so short that come up with too many alignments which is hard to filter out correct result. Thus, this software is mainly for RNA-Seq reads longer than 70 bp.

BAsplice is an effective tool on align NGS to reference, and it also can detect both nonconservative and conservative splice junctions. It can provide information for doing research on RNA-Seq. Compare with other existing tool (SOAPsplice), BAsplice have a better performance under different read lengths and sequencing depth.

IV. ACKNOWLEDGMENT

This study was supported by grant from the Special Foundation Work Program (No. 2009FY120100), the Ministry of Science and Technology of the People's Republic of China; grant from the National Science Foundation of China (No. 31071163 and 31101063).

V. REFERENCES

[1] Wang Z, Gerstein M, Snyder M.,"RNA-Seq: a revolutionary tool for transcriptomicsrevolutionary tool for transcriptomics", Nat Rev Genet. 2009 Jan;10(1):57-63.

[2] Green,M.R. " Pre-mRNA splicing". Annu. Rev. Genet. 1986, 20, 671–708.

[3] Li, H. and Durbin, R., Fast and accurate short read alignment with BurrowsWheeler transform. Bioinformatics, 2009, 25(14), 1754–1760.

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

[5] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B,"Mapping and quantifying mammalian transcriptomes by RNA-Seq", Nat Methods. 2008 Jul;5(7):621-8

[6] De Bona F, Ossowski S, Schneeberger K, Rätsch G."Optimal spliced alignments of short sequence reads", Bioinformatics. 2008 Aug 15;24(16):i174-80

[7] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. Bioinformatics, 25(9), 1105–1111.

[8] 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. "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery", Nucleic acids research, 2010, 38(18), e178.

[9] Au, Kin Fai, Jiang, Hui, Lin, Lan, Xing, Yi, Wong, Wing Hung. "Detection of splice junctions from paired-end RNA-seq data by SpliceMap." Nucleic Acids Research, 2010, 38(14), 4570-4578.

[10] Huang S, Zhang J, Li R, Zhang W, He Z, Lam T-W, Peng Z and Yiu S-M, "SOAPsplice: genome-wide ab initiodetection of splice junctions from RNA-Seq data". Front. Gene, 2011, 2:46.

[11] Zhong Wang, Mark Gerstein, and Michael Snyder,"computational methods for transcriptome annotation and quantification using RNA-seq", Nature Methods, 2011, 8,469

[12] Li H, Ruan J, Durbin R. "Mapping short dna sequencing reads and calling variants using mapping quality scores", Genome Res. 2008, 18, 1851–1858.

[13] Jiang H, Wong WH. "SeqMap: mapping massive amount of oligonucleotides to the genome", Bioinformatics. 2008 Oct 15;24(20):2395-6.

[14] ELAND: http://bioinfo.cgrb.oregonstate.edu/docs/solexa/

[15] Ruiqiang Li, Tam A., Wong, S., Wu, E., Yiu, S.M., "High Throughput Short Read Alignment via Bi-directional BWT",2009, IEEE International Conference on Bioinformatics and Biomedicine.

[16] Ferragina,P. and Manzini,G. "Opportunistic data structures with applications". In Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000), IEEE Computer Society, pp. 390–398.

[17] wgsim: https://github.com/lh3/wgsim

[18] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup(2009). The sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE