2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

A Comparative Analysis System for Detecting Alternative Splicing in Multiple Ngs Result Files

Sora Kim , Seokmoon Choi and Hwan-Gue Cho

Dept. of Computer Science and Engineering, Pusan National University, Busan, Korea

Abstract. Next generation sequencing technology enables massive high-throughput mRNA sequencing(RNA-seq). RNA-seq is useful for finding alternative splicing, novel genes, and structural variations. There are a number of splice junction detecting tools; however, the output of each of them is slightly different from the others. As a result, scientists confuse outcomes what is true or where is intersection. Here we introduce the common junction format (CJF). This format shows where common junction sites are and where alternative splicings are suspected. It is created by joining results from the splice junction detecting tools. TheCJF format supports the BED format and the JUNC format. The BED format is output from TopHat, SpliceMap, MapSplice, andHMMsplicer whereas the JUNC format is output from SOAPsplice. To test the CJF format, we took the outcome from each tool and made a CJF file from each outcome. We then detected each tool's alternative splicing and junction estimation. And we also designed CJF visualizer to show the result.

Keywords: Alternative splicing, RNA-seq, CJF, BED format

1. Introduction

In recent studies, RNA-seq has primarily been used to quantify the expression levels of specific tissue and genes[1,2]. RNA-seq, a recently developed approach to transcriptome profiling, uses deep-sequencing technologies[3].

Alternative splicing is known to affect more than half of all human genes and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals[4].Early studies of alternative splicing events were based on EST (Expressed Sequence Tag) libraries[5]. However, with the emergence of NGS technologies, RNA-seq has been introduced as a tool for the study of alternative splicing[6,7,8],and many novel alternative splicing events have been detected using its data[9].

There are many tools used to detect splice junctionand make conclusions about them. However, their outcomes are slightly different because each tool used its own algorithm. As a result, biologists commonly confuse the results provided by the different tools since theycannotdirectly compare them. Here we introduce a new file format, the common junction format (CJF), in an effort to solve this problem.

2. Common junction format

TheCJF format is used to compare different tools' junction sites. Table 1 displays the CJF attribute information. In other words, tools that detect splice junctions such as TopHat[10], SpliceMap[11], MapSplice[12], HMMSplicer[13], and SOAPsplice[14]have differing splice junction site results. We created CJF format file as a result. The process of creating a CJF file is shown below.

2.1. Make a data file

We first obtained the result files from the tools used to detectsplice junction. TopHat, SpliceMap, MapSplice, and HMMsplicercreated results in the BED format. SOAPsplicecreated results as s JUNC format. The JUNC format is only used in SOAPsplice. The JUNC format attributes consist of {chrom, junctionStart,

junctionEnd, strand, quality}. “chrom” is the chromosomal information that is found the junction site,“junctionStart” is the junction site start position in “chrom”,and “junctionEnd” is the junction site end position in “chrom”.

2.2. Junction lists extraction

We obtained the junction lists from the tool outcome. If the outcome format was BED, we calculated the junction site using the block position. If the outcome format was JUNC, we simply extracted the junction site. We then made a list ofjunction sites.

2.3. Common junction list creation

Using a list about junction sites, we made a common junction listthat we called CJFconvertor. Using a junction start position, theCJFconvertor first sorts a junction site listand then compares the list’s first and second items. If they are same, CJFconvertor mergesthese items and creates a common junction list. In other words, CJFconvertor makes a common junction list using the first and second items.

2.4. Find alternative splicing

When CJFconvertormakes a common junction list, it also checks for alternative splicing. Figure 1 shows the junction lists extracted by theCJFconvertor.The green sticks are junctions. The red boxes are exons. Junctions j_1 and j_2 are overlapped, meaning that junction j_1 has an exon site and that is alternatively spliced. Thus, we predict that junctions j_1 and j_2 are both alternative spliced.

Table 1 shows the CJF format’s attribute. “chrom” is the chromosome number(e.g. chr1 and chr2) on which the junction is located. Humans have 24 chromosomes including chrX and chrY. However, if chromosome 1 only appears at the junction in the experiments, then the CJF result shows only chr1’s junction site list. “junctionStart” and “junctionEnd” are the junction site positions. “densityNumber” is the number of tools finding the junction site. If junction is found by TopHat, SpliceMap, andMapSplice, then its densityNumber is 3. “density” is theproportion of the tools finding the junction site. For example, in this experiment, we used 5 tools,but only 3 tools found junction . As a result, junction ’s density is 60%(3/5). “tools” is a list of tools finding the junction site. “alternative” means alternative splicing appearance.

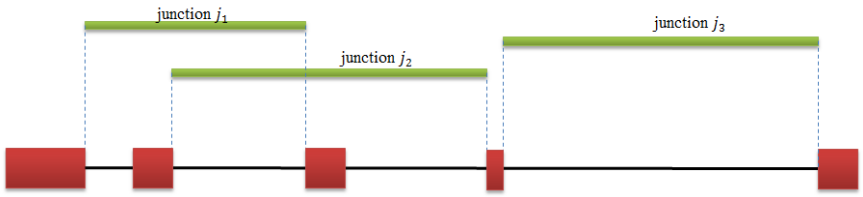

Fig.1 Green sticks are junctions. Red boxes are exons. Junction lists are extracted from the splice junction detecting tool's outcome.

Tab.1Common junction format’s attribute

Num.	Field Attribute	Information
1	chrom	Chromosome location that found the junction site (e.g. chr1, chr2, …, chrY)
2	junctionStart	Junction site start position in chromosome
3	junctionEnd	Junction site end position in chromosome
4	densityNumber	Number of tools finding junction site
5	density	Proportion (%) of tools finding a junction site. N(tools finding junction site)/N(total tools)
6	tools	Names of tools finding the junction site
7	alternative	Alternative splicing’s appearance or not

3. Results

We used theTopHat, SpliceMap, MapSplice, and SOAPsplicetools in the experiment. HMMsplicer is usable only for single-end read data. However, our simulation data are paired-end read data.

Figure 2 shows the visualizer resultsof the CJF format. Number 1 is the gene name by searching. Number 2 is the transcript's name. Number 3 is the known exon block. Numbers 4 and 5 are explained in Figure 3. Figure 3 shows the junction (A) and alternative splicing (B). Number 6 is the read depth mapped on the whole genome. Number 7 is the length of the reference genome

Figure 4 shows the result of searching several genes. Numbered areas 1-3 show each gene's exon and junction results. We checked the CJF format result alternative splicing appearance and the visualizer's view and verified that they corresponded 100%.

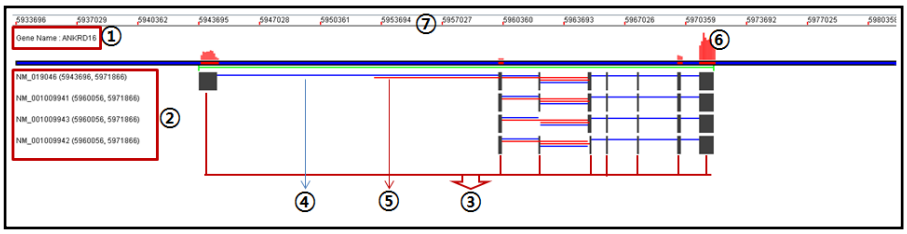

Fig.2 Common junction format results

Fig.3 Junction (A) and alternative splicing (B) of numbered areas
4 and 5, respectively, of Figure 2

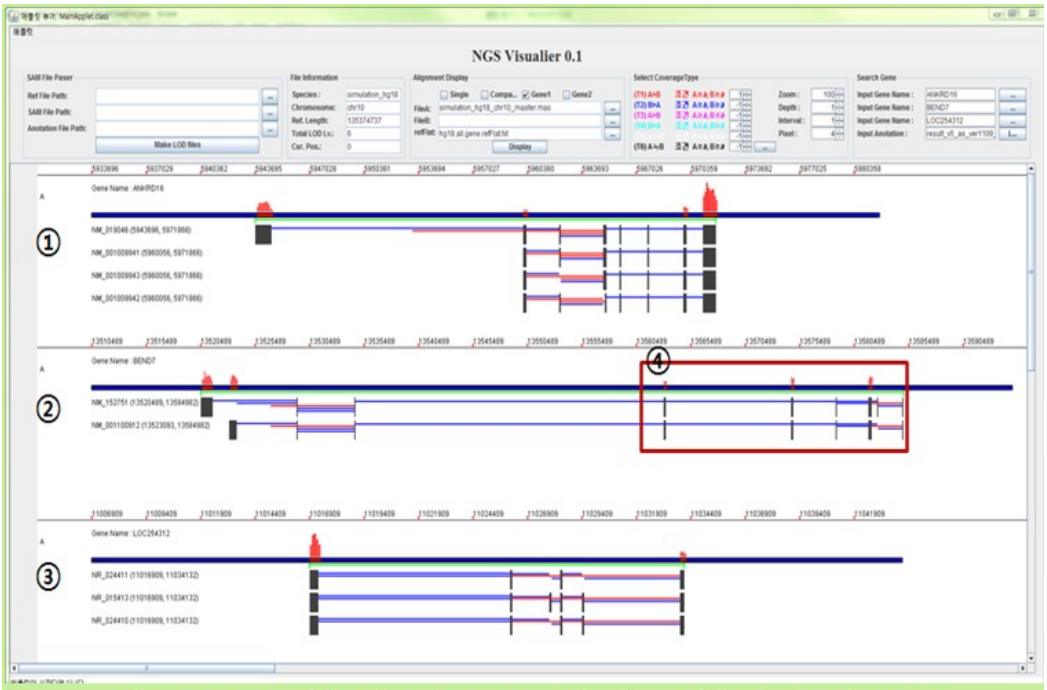

Fig.4 Common junction format result of searching several genes

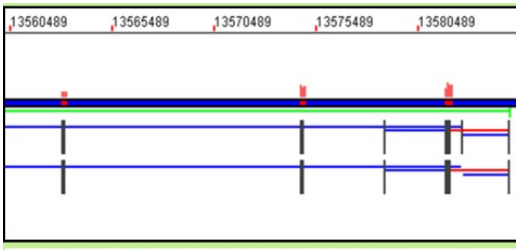

Fig.5 Enlarged view of box 4 in Figure 4

Figure 2 shows the visualizer resultsof the CJF format. Number 1 is the gene name by searching. Number 2 is the transcript's name. Number 3 is the known exon block. Numbers 4 and 5 are explained in Figure 3. Figure 3 shows the junction (A) and alternative splicing (B). Number 6 is the read depth mapped on the whole genome. Number 7 is the length of the reference genome.

Figure 4 shows the result of searching several genes. Numbered areas 1-3 show each gene's exon and junction results. We checked the CJF format result alternative splicing appearance and the visualizer's view and verified that they corresponded 100%.

4. Conclusions

RNA-seq technology offers ability to accurately measure transcript abundances in an RNA sample [15]. Unfortunately, current technological limitations of sequencers result incDNA molecules representing only partial fragments of the RNA being probed[16].

Due to increasing interest in alternative splicing, many tools are being developed. However, the different tools' outcomes are not always the same. As a result, we created the CJF, results of which confirm that alternative splicing results exactly match those of the real visualizer view.

In future studies, we will update the CJF format with regard to alternative splicing type, which will allow us to check the file and determine the junction's alternative splicing type.

5. Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2011-0015359)

6. References

[1] Fatih O., Patrice M. M. RNA sequencing: advances, challenges and opportunities, Nat. Rev. Genet., 2011.12, PP87-98

[2] France D., Jean-Marc A., Corinne D. S., et al. Annotating genomes with massive-scale RNA sequencing, Gen. Biology, 2008.9, R175

[3] Zhong W., Mark G., Michael S. RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet., 2009.10, PP57-63

[4] Eric T. W., Rickard S., Shujun L., et al. Alternative isoform regulation in human tissue transcriptomes, Nature, 2008.456, PP470-476

[5] Mark D. A., Anthony R. K., Chris F., J. Craig V. 3,400 new expressed sequence tags identify diversity of transcripts in human brain, Nat. Genet., 1993.4, PP256-267

[6] Qun P., Ofer S., Leo J. L., et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet., 2008.40, PP1413-1415

[7] Barmak M., Alissa R., Catherine G., Christopher L. Genome-wide detection of alternative splicing in expressed sequences of human genes, Genome Res., 2010.20, PP45-58

[8] Ali M., Brian A. W., Kenneth M., et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods, 2008.5, PP621-628

[9] Cole T., Brian A. W., Geo P., et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechno., 2010.28, PP511-515

[10] Trapnell, C., Pachter, L., Salzberg, S.L. TopHat: discovering splice junctions with RNA-seq, Bioinformatics, 2009.25, PP1105-1111

[11] Au, K. F., Jiang, H., Lin, L., et al.. Detection of splice junctions from paired-end RNA-seq data by SpliceMap, Nucleic Acids Res, 2010.38, PP4570-4578

[12] Wang, K., Singh, D., Zeng, Z., et al..MapSplice: accurate mapping of RNA-seq reads for splice junction discovery, Nucleic Acids Res, 2010.38, PPe178

[13] Dimon, M. T., Sorber, K., DeRisi, J. L. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-seq data. PLoS ONE, 2010.5, PPe13875

[14] Songbo H., Jinbo Z., Ruiqiang L., et al..SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data, front. In GENETICS, 2011.2, PP1-12

[15] Marguerat S., Bahler J. RNA-Seq: from technology to biology, Cellular and Molecular Life Sciences, 2010.67, PP569-579

[16] Adam R., Cole T., Julie D., et al. Improving RNA-Seq expression estimates by correcting for fragment bias, Gen. Biology, 2011.12, R22