BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1708–1710
doi:10.1093/bioinformatics/btr265

Sequence analysis

Advance Access publication May 5, 2011

FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq

Yang Li[1], Jeremy Chien[3], David I. Smith[3] and Jian Ma[1,2,*]

[1]Department of Bioengineering, [2]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 and [3]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Fusion transcripts can be created as a result of genome rearrangement in cancer. Some of them play important roles in carcinogenesis, and can serve as diagnostic and therapeutic targets. With more and more cancer genomes being sequenced by next-generation sequencing technologies, we believe an efficient tool for reliably identifying fusion transcripts will be desirable for many groups.

Results: We designed and implemented an open-source software tool, called FusionHunter, which reliably identifies fusion transcripts from transcriptional analysis of paired-end RNA-seq. We show that FusionHunter can accurately detect fusions that were previously confirmed by RT-PCR in a publicly available dataset. The purpose of FusionHunter is to identify potential fusions with high sensitivity and specificity and to guide further functional validation in the laboratory.

Availability: http://bioen-compbio.bioen.illinois.edu/FusionHunter/.

Contact: jianma@illinois.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 19, 2011; revised on March 16, 2011; accepted on April 13, 2011

1 INTRODUCTION

One of the key features observed when analyzing cancer genomes is the chromosomal abnormality. Genome arrangements could result in aberrant fusion genes, and a number of them have been found to play important roles in carcinogenesis. Different functional fusion genes have been detected in both hematological malignancies and solid tumors (Mitelman *et al.*, 2007). Precisely identifying these fusion genes is important for developing potential diagnostic and therapeutic targets.

Recently, RNA-seq has proven to be a valuable tool for transcriptome profiling based upon ultra high-throughput next-generation sequencing technologies. It has been shown that RNA-seq is a more accurate method to survey the entire transcriptome in a quantitative and high-throughput fashion than microarray technology and EST sequencing (Wang *et al.*, 2009). RNA-seq can provide a great deal of information about genome wide transcription, including the identification of fusion genes in cancer. A number of novel fusion genes has been identified using single-end RNA-seq reads (Maher *et al.*, 2009b) and paired-end RNA-seq reads (Berger *et al.*, 2010; Maher *et al.*, 2009a). The advantage of paired-end technology is that it provides a more reliable way to uncover breakpoints when comparing two genomes.

However, although a computational pipeline that identifies fusion transcripts using RNA-seq has been mentioned in a number of recently published cancer transcriptomic studies (Berger *et al.*, 2010; Maher *et al.*, 2009a,b), there was no publicly accessible, open-source software available until a recently published tool named FusionSeq (Sboner *et al.*, 2010). With more and more cancer transcriptomes being sequenced by next-generation sequencing, we believe an efficient tool for reliably identifying fusion transcripts will be desirable for many groups. Here, we describe an open-source software tool, called FusionHunter, that identifies fusion transcripts in cancer, using reads from paired-end RNA-seq. This high-throughput software tool is designed for efficient and precise detection of fusion transcripts in cancer for further functional validation in the laboratory.

2 METHODS

(i) Map the RNA-seq reads to the reference genome: currently, the input of FusionHunter should be paired-end RNA-seq reads. These original reads (e.g. 50mers) are uniquely mapped to the reference human genome using Bowtie (Langmead *et al.*, 2009). One strength of our tool is that it has the option to define transcript fragments without relying on known annotations in order to find novel transcripts [see Fig. 1A]: we first identify potential exons by clustering reads together; and then we further group exons into potential transcript fragments. However, in our implementation, we include the annotation from UCSC known genes to make the results more reliable.
(ii) Find candidate fusions: we call a candidate fusion if two transcripts are both enriched by input RNA-seq reads and the linkage between them is supported (encompassed) by at least two independent paired-end reads. However, even though we require unique mapping for the original RNA-seq reads, there will still be unreliable mappings that may introduce false positives. This is mainly due to duplications in the human genome. Figure 1B illustrates such a situation. We have a read pair with reads F and R. The correct mapping of F and R should be located in A. But R could be mapped to A's homolog B, possibly due to sequencing errors on the 3′ end of R. This will mistakenly link A and B together as a candidate fusion transcript. To resolve this problem, we look at the self-alignment of the human genome and the logic is similar to the idea of excluding fusions between homologous regions in Hu *et al.* (2010). If a candidate fusion transcript consists of two genes that share significant homology, it will be removed. After this filtering, we create a 'pseudo reference' for each candidate. Briefly, if we have paired-end reads supporting a fusion between genes A and B, we first determine the relative order of them (i.e. A B or A -B or -A B or -A -B). Then, we make a 'pseudo reference' by concatenating sequences of A and B according to their relative order (e.g. if the order is A -B, we concatenate the reverse complement of B

*To whom correspondence should be addressed.

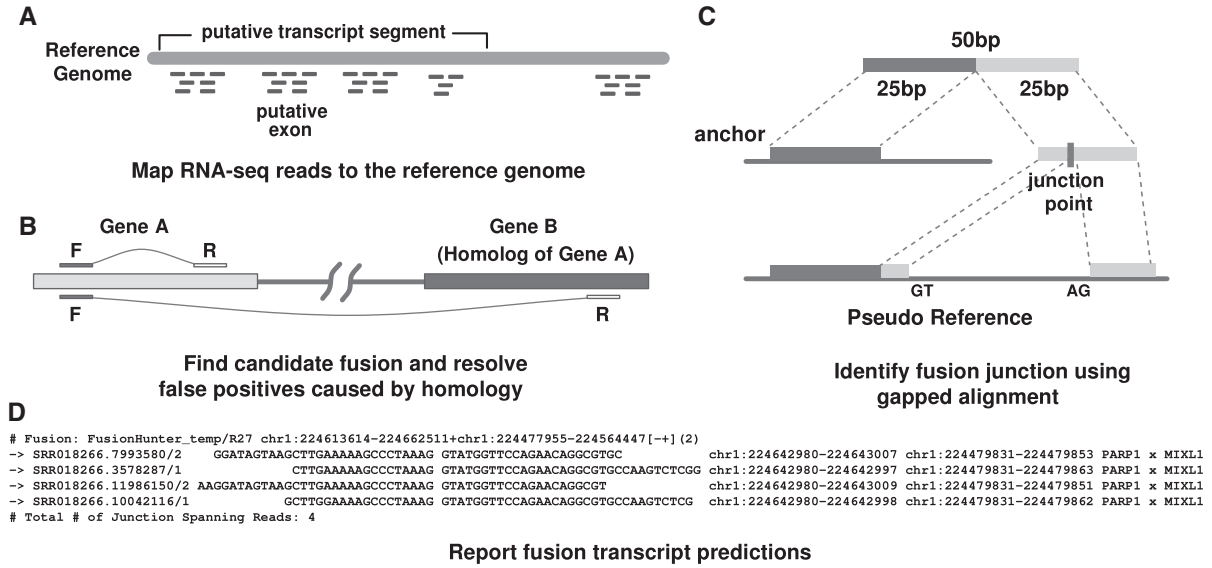

Fig. 1. Key steps in FusionHunter. (**A**) Reads are mapped to the reference human genome to detect putative exons and putative transcripts. (**B**) Encompassing reads are used to select candidate fusions. Special attention is paid to resolve false positives caused by homology. (**C**) A gapped alignment implemented in FusionHunter to identify exact fusion junctions. (**D**) The report of a fusion gene PARP1-MIXL1 in 501-MEL. The number of encompassing reads, fusion junction spanning reads and co-ordinates in the reference human genome are displayed.

to the end of A). We create such a pseudo reference for each candidate. All the subsequent analysis on these pseudo references can be run in parallel to make it more efficient.

(iii) Identify fusion junction spanning reads: based on the pseudo reference of each candidate region, we also want to find RNA-seq reads that precisely span the fusion junction. This serves as another criteria to reduce false positives from the previous step and the junction spanning reads provide strong support for fusion events. We implemented a gapped alignment method similar to the algorithms described by previously published works (Au *et al.*, 2010; Wang *et al.*, 2010) to detect fusion spanning reads. For a 50 bp read, which cannot be aligned to the pseudo reference, FusionHunter splits it into two 25 bp segments. One of these segments, which we call the 'anchor', would have an alignment on the pseudo reference if the original read is junction spanning. Currently, we assume that each read can span two exons at most. As shown in Figure 1C, FusionHunter then tries to extend the anchor till it reaches a canonical splicing signal (Burset *et al*., 2000), and search alignment for the remainder of the original read on the pseudo reference. Since we will obtain many candidate regions from the previous step, this gapped alignment procedure utilizes a multi-core approach to speed up, where the number of total threads could be configured by the users. The output of the gapped alignment is in the SAM format.

(iv) Report fusion transcripts: based on the SAM file produced in the previous step, an RNA-seq read is categorized as a fusion spanning read if it is split (referring to the 'CIGAR' column of the SAM file) into two alignments to different genes in the candidate region (pseudo reference). A fusion is reported if at least one read spanning the fusion boundary has been found. To reduce false positives, we discard junction spanning reads with less than 6 bp matches on either gene. We remove potential PCR artifacts by retaining only one read for each mapping co-ordinate on the reference, and we require candidate fusions that are supported by only one junction spanning read to originate from annotated splice sites. For each detected fusion, names and co-ordinates of both genes in the human reference (hg18), numbers of fusion encompassing and spanning paired-end RNA-seq reads, are all represented in the final output. Sequences of reads spanning fusions are presented with a gap where the fusion junction lies. In addition to reporting fusions, FusionHunter also reports read-through events, in which two nearby genes (within 600 kb and separated by at most one gene) are co-transcribed in the same orientation. Since many of the read-through events have already annotated by human Expressed Sequence Tags (EST), FusionHunter only reports novel read-throughs that are not shown in human EST database.

3 RESULTS

Implementation and running time: major components in FusionHunter were implemented in C and C++. Perl scripts were used to wrap different parts into a pipeline. In the current version of FusionHunter, we require the users to install Bowtie which is freely available. For a sample of 30 million reads, it takes 1–2 h to align reads using Bowtie and 0.5–1 h for downstream processes, using computers with 16 cores.

Au *et al. (2010) datasets*: we ran FusionHunter on the RNA-seq sample from normal human brain tissue from Au *et al.* (2010), which serves as a control. FusionHunter did not report any fusion events.

Berger *et al. (2010) datasets*: Berger *et al*. (2010) validated 11 novel gene fusions through RT-PCR from ten melanoma samples. We downloaded the raw RNA-seq reads of these samples used in their study from NCBI and ran FusionHunter on these datasets. FusionHunter predicted 11 fusions and all of them were validated by Berger *et al.* (2010). This shows that fusion predictions of FusionHunter are reliable. Details of the results are in the Supplement. In addition, we also predicted 13 read-through events and five of them overlapped with 12 read-throughs predicted in Berger *et al.* (2010). We did not perform detailed comparison for read-throughs, because Berger *et al.* (2010) did not validate the read-through predictions.

4 DISCUSSION

Recently, several works have been published to detect splice junctions and chimeric transcripts based on RNA-seq, including

FusionSeq, MapSplice (Wang *et al*., 2010) and SplitSeek (Ameur *et al*., 2010). FusionSeq reports fusion events as well as their quality scores. The methodology of how FusionSeq detects fusion junctions is different from FusionHunter. FusionSeq splits exonic sequences from two candidate fusion genes into a large number of short 'tiles' and merges all possible pairs to create a 'fusion junction library'. Then, it tries to align candidate fusion junction reads against the junction library using Bowtie. This process is less efficient since the fusion junction library is normally quite large. Also, FusionSeq considers only annotated exons and may miss events on novel exons. MapSplice and SplitSeek are tools designed to align RNA-seq reads to splice junctions. They can potentially be used to identify long-range chimeric splicing events, although the tools were not designed for efficient and reliable detection of fusion transcripts. There is also a very recent method that aims to detect fusions more sensitively utilizing ambiguously mapping of RNA-seq read pairs (Kinsella *et al*., 2011). A number of improvements can be employed to further aid in the identification of true fusion genes. For example, by comparing the breakpoints identified from RNA-seq reads and genomic DNA reads, we will be able to know whether the fusion genes detected in the transcriptome are caused by rearrangements at the genomic level or by other mechanisms (e.g. *trans*-splicing). Future versions of FusionHunter will accomplish this by incorporating paired-end/mate-pair reads from whole-genome DNA sequencing.

ACKNOWLEDGEMENT

We thank Jaebum Kim for testing the software.

Funding: National Center for Supercomputing Applications (Faculty fellowship); the Mayo-UIUC Alliance (Planning grant).

Conflict of Interest: none declared.

REFERENCES

Ameur,A. *et al*. (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.*, **11**, R34.

Au,K.F. *et al.* (2010) Detection of splice junctions from paired-end RNAseq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.

Berger,M.F. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.*, **20**, 413–427.

Burset,M. *et al.* (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.

Hu,Y. *et al.* (2010) A probabilistic framework for aligning paired-end RNA-seq data. *Bioinformatics*, **26**, 1950–1957.

Kinsella,M. *et al*. (2011). Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, **27**, 1068–1075.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Maher,C.A. *et al.* (2009a) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

Maher,C.A. *et al*. (2009b) Transcriptome sequencing to detect gene fusions in cancer. *Nature,* **458**, 97–101.

Mitelman,F. *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer,* **7(4)**, 233–245.

Sboner,A. *et al*. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.,* **11**, R104.

Wang,K. *et al*. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Wang,Z. *et al.* (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10(1)**, 57–63.