BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 14 2012, pages 1923–1924
doi:10.1093/bioinformatics/bts272

Sequence analysis

BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data

Ken Chen[1,*], John W. Wallis[2,3], Cyriac Kandoth[2], Joelle M. Kalicki-Veizer[2], Karen L. Mungall[4], Andrew J. Mungall[4], Steven J. Jones[4], Marco A. Marra[4], Timothy J. Ley[2,5], Elaine R. Mardis[2,3], Richard K. Wilson[2,3], John N. Weinstein[1] and Li Ding[2,3,*]

[1]Department of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center, Houston, TX, [2]The Genome Institute, [3]Department of Genetics, Washington University, St Louis, MO, USA, [4]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada and [5]Department of Internal Medicine, Division of Oncology, Washington University, St Louis, MO, USA

Associate Editor: Michael Brudno

ABSTRACT

Summary: Despite recent progress, computational tools that identify gene fusions from next-generation whole transcriptome sequencing data are often limited in accuracy and scalability. Here, we present a software package, BreakFusion that combines the strength of reference alignment followed by read-pair analysis and *de novo* assembly to achieve a good balance in sensitivity, specificity and computational efficiency.

Availability: http://bioinformatics.mdanderson.org/main/BreakFusion

Contact: kchen3@mdanderson.org; lding@genome.wustl.edu

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

Received on January 9, 2012; revised on March 8, 2012; accepted on May 1, 2012

1 INTRODUCTION

Despite recent progress, bioinformatics tools that analyze RNA-seq data are often limited in accuracy and scalability. Alignment-based tools (Garber *et al*., 2011) typically predict such large numbers of candidates that further examination of the data is impractical. Furthermore, many findings are artifacts derived from a large number of erroneous short reads produced by the next-generation sequencing (NGS) instruments or are read misalignments produced by the sequence aligners. Assembly-based approaches (Martin and Wang, 2011) can potentially achieve higher accuracy because they leverage dependency among reads and are less sensitive to errors in individual reads. With sufficient coverage, longer and higher quality sequences can be assembled from short reads, resulting in improved specificity. However, existing assembly-based approaches (Garber *et al*., 2011; Grabherr *et al*., 2011; Robertson *et al*., 2010; Trapnell *et al*., 2010) potentially lack sensitivity to rare events such as gene fusions because they are designed to reconstruct entire transcriptomes, rather than focus on novel sequences. We reason that a targeted transcriptome assembly strategy that focuses on assembling novel junction sequences can potentially achieve a better balance in sensitivity, specificity and computational efficiency.

*To whom correspondence should be addressed.

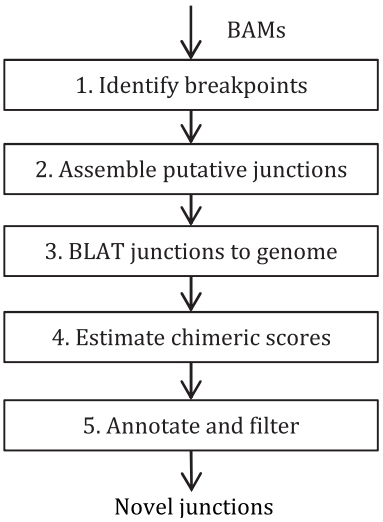

Fig. 1. BreakFusion overview

2 METHODS

Here we present an NGS Bioinformatics pipeline, BreakFusion that implements a targeted assembly approach for novel transcriptomic sequence discovery.

BreakFusion consists of five steps, as illustrated in Figure 1. The input is one or a set of whole transcriptome BAM files that contains mapped paired-end RNA-seq reads. The mapping can be performed using a set of algorithms such as BWA and TopHat, preferably including known splice junctions in the references to achieve good alignment. Step 1 identifies splicing breakpoints using a read-pair algorithm such as BreakDancer (Chen *et al*., 2009) or any splice mapping algorithm such as TopHat-Fusion (Kim and Salzberg, 2011). Step 2 locally assembles short reads that are anchored around each breakpoint using TIGRA and produces a set of splice junction contigs that are supported by both mapped and one-end-anchored reads (Mills *et al*., 2011). TIGRA is a *de Bruijn* graph-based assembler that is specially designed to construct all possible allelic sequences at loci of heterogeneous origin. It is sufficiently sensitive to identify low abundance alleles as long as their constituent *k*mers are observed at least twice in the reads. Step 3 aligns the junction sequences to the genome using BLAT (Kent, 2002). Step 4 summarizes BLAT alignments into a single chimeric score that quantifies the likelihood of an

assembled junction sequence containing *bona fide* breakpoints relative to the genome. Step 5 annotates the breakpoints using UCSC databases and filters the breakpoints by mainly two factors: (i) the chimeric scores computed in Step 4, and (ii) the self-chain alignment annotation, which indicates whether the breakpoints are caused by misalignment in duplicated regions.

Most steps in the BreakFusion pipeline are performed by previously published and well-attested algorithms. The chimeric scoring system in Step 4 is novel to this work. By default, BLAT does not produce a single hit for transcripts that span two chromosomes or distances >750 000 bp. Instead, it reports individual alignments for each of the sub-segments, with no direct indication of the existence of the chimeric structure and the level of confidence associated with it. To overcome such limitation, we first remove hits that are not unique (i.e. map to multiple regions). We then chain the remaining alignments into longer ones if they can form 1-monotonic maps (Brudno *et al*., 2003). We compute scores for chained alignment using the same BLAT formula that is used for each constituent alignment. We identify the best and the second best alignment after chaining, and compute a chimeric score using the following equation:

$$c = e^{\frac{(q^* - q^0)}{10}} - e^{\frac{(q' - q^0)}{10}},$$

Where q^* is the alignment score of the best chained alignment, q' the second best, and q^0 is the length of the query sequence. This equation produces scores between 0 and 1.0. The score becomes high when the best alignment well explains the entire query sequence (q^* approaches q^0) and the second best alignment is appreciably worse ($q' \ll q^*$). The constant 10 makes the scores sensitive to differences at the 10-bp scale.

3 RESULTS

To test BreakFusion's performance, we compared it with TopHat-Fusion and defuse (McPherson *et al*., 2011) using four publicly available RNA-seq datasets: a prostate cancer cell line (NCIH660) and a matched lymphoblastoid cell line (GM12878) (Sboner *et al*., 2010), a breast cancer cell line (MCF-7) (Edgren *et al*., 2011) and a chronic myelogenous leukemia cell line (K-562) (Berger *et al*., 2010). BreakFusion achieved better sensitivity and specificity tradeoff in these tests (Supplementary Data). It also used substantially less CPU time. We further applied BreakFusion to analyze RNA-seq data for 155 acute myeloid leukemia (AML) samples as part of The Cancer Genome Atlas (TCGA) project. This dataset was also analyzed by the Genome Sciences Centre at the BC Cancer Agency (BCCA) using Tran-Abyss (Robertson *et al*., 2010), which resulted in 67 instances of experimentally validated fusions. BreakFusion was able to rediscover all these fusions and additionally predicted eight new instances of fusions, which were subsequently confirmed by the BCCA group (Supplementary Data).

4 DISCUSSION

To the best of our knowledge, BreakFusion is the first approach that performs targeted assembly on RNA-seq data for fusion identification. The results of our experiments indicate that BreakFusion has achieved sensitivity and specificity comparable or better than other tools, and is clearly more computationally efficient. This, in our view, represents a methodology improvement that will benefit many projects that utilize NGS RNA sequencing data. Besides fusion discovery, in principle, BreakFusion can be applied to identify novel alternative splicing events.

The component programs of BreakFusion are efficiently implemented in C++ and perl. It finished analyzing 155 TCGA AML BAM files in <2 h using 155 CPUs with 4 GB RAM each. BreakFusion is freely available for academic use at http://bioinformatics.mdanderson.org/main/BreakFusion.

ACKNOWLEDGEMENTS

We thank Hao Zhao, Jianping Zhang, Heather Schmidt and Christopher Miller for their assistance. We are also grateful to Nina Thiessen, Elizabeth Chun and Gordon Robertson from the BCCA, as well as the members of TCGA AML Analysis Working Group.

Funding: The Cancer Center Support Grant from NCI (P30 CA016672 to UT MD Anderson); Center for Large-Scale Genome Sequencing and Analysis from NHGRI (U54 HG003079 to R.K.W); Structural Genomic Variation analysis for the 1000 Genome project from NHGRI (U01 HG005209 to L.C.)

Conflict of Interest: none declared.

REFERENCES

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

Brudno,M. *et al.* (2003) Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, **19** (Suppl. 1), i54–i62.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, **6**, 677–681.

Edgren,H. *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing, *Genome Biol.*, **12**, R6.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Kent,W.J. (2002) BLAT - The BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kim,D. and Salzberg,S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

McPherson,A. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Robertson,G. *et al.* (2010) De novo assembly and analysis of RNA-seq data, *Nat Methods*, **7**, 909-U962.

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

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.