Sequence analysis

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

Ken Chen¹,*, John W. Wallis²,³, Cyriac Kandoth², Joelle M. Kalicki-Veizer², Karen L. Mungall⁴, Andrew J. Mungall⁴, Steven J. Jones⁴, Marco A. Marra⁴, Timothy J. Ley⁵,⁶, Elaine R. Mardis³,³, Richard K. Wilson²,³, John N. Weinstein¹ and Li Ding²,³,³

¹Department of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center, Houston, TX, ²The Genome Institute, ³Department of Genetics, Washington University, St Louis, MO, USA, ⁴Canada’s Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada and ⁵Department 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 (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 (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 (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 (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 (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 (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 (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 (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 (Martin and Wang, 2011) can potentially achieve higher accuracy because they leverag
K. Chen et al.

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 chromatic 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 chromic 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 chromic 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-motonic 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 chromic score using the following equation:

$$s = \frac{c}{q * 10} - \frac{q}{q'}$$

Where \(q\) is the alignment score of the best chained alignment, \(q'\) the second best, and \(q\) 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 \(q\)) and the second best alignment is appreciably worse (\(q' < q\)). 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 (NCI-H660) and a matched lymphoblastoid cell line (GM12878) (Shomer 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). 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