Bioinformatics, 32(8), 2016, 1226–1228
doi: 10.1093/bioinformatics/btv721
Advance Access Publication Date: 10 December 2015
Applications Note

Sequence analysis

FuMa: reporting overlap in RNA-seq detected fusion genes

Youri Hoogstrate[1,2], René Böttcher[1], Saskia Hiltemann[1,2], Peter J. van der Spek[2], Guido Jenster[1] and Andrew P. Stubbs[2,*]

[1]Department of Urology and [2]Department of Bioinformatics, Erasmus University Medical Center, Rotterdam, 3000 CA, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 25 June 2015; revised on 30 November 2015; accepted on 4 December 2015

Abstract

Summary: A new generation of tools that identify fusion genes in RNA-seq data is limited in either sensitivity and or specificity. To allow further downstream analysis and to estimate performance, predicted fusion genes from different tools have to be compared. However, the transcriptomic context complicates genomic location-based matching. *FusionMatcher* (FuMa) is a program that reports identical fusion genes based on gene-name annotations. FuMa automatically compares and summarizes all combinations of two or more datasets in a single run, without additional programming necessary. FuMa uses one gene annotation, avoiding mismatches caused by tool-specific gene annotations. FuMa matches 10% more fusion genes compared with exact gene matching due to overlapping genes and accepts intermediate output files that allow a stepwise analysis of corresponding tools.

Availability and implementation: The code is available at: https://github.com/ErasmusMC-Bioinformatics/fuma and available for Galaxy in the tool sheds and directly accessible at https://bioinf-galaxian.erasmusmc.nl/galaxy/

Contact: y.hoogstrate@erasmusmc.nl or a.stubbs@erasmusmc.nl

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

1 Introduction

A new generation of bioinformatics tools has been released that aims to detect fusion genes within RNA-seq data; however, the current tools are limited in either sensitivity or specificity, making their results impractical for downstream analysis and subsequent validation (Carrara *et al.*, 2013). As shown in other domains of high throughput sequencing analysis, using a consensus of tools may improve performance by compensating for individual tool error profiles (Ewing *et al.*, 2015; Goode *et al.*, 2013). Since no single tool shows superior detection performance, a consensus-based fusion gene detection in RNA-seq can improve both downstream analysis as well as overall performance. Moreover, to identify limiting factors and promote improvement of current algorithms, an accurate estimation of sensitivity and specificity is required, for which an accurate fusion gene comparison is crucial. Therefore, comparing validated and *in silico* predicted fusion genes in an automated fashion and summarizing identical fusion genes in two or more datasets in an easy and accessible way is a desirable feature.

Ideally, sensitivity/specificity estimation should be based upon the identified genomic breakpoints provided as two chromosomal locations. However, the nature of RNA-seq complicates this strategy, as reads may span exon junctions and breakpoints are more likely to be expected in introns because of their relative large size, introducing additional uncertainty when trying to pinpoint the exact genomic position of the DNA breakpoint. For instance, using exact position-based matching (EPM) results in poor overlap between tools (Supplementary Section S6). A less conservative strategy used in DNA-seq analysis involves comparing genomic intervals (Layer *et al.*, 2014), which can be defined by adding flanking regions to each breakpoint to increase the likelihood of matching fusion events called by multiple programs. In RNA-seq analysis this strategy is also not sufficient, because the intervals are not related to the organization of the transcriptome. In addition, transcriptome annotations differ substantially between sources which frequently result in inconsistent results in RNA-seq analysis (Zhao and Zhang, 2015), intron sizes are differing and lastly, a substantial number of genes do overlap with each other (Sanna *et al.*, 2008) due to opposite strand positioning. Therefore, we designed a new method, FuMa, which boosts the current fusion gene matching functionality of the Chimera package (Beccuti *et al.*, 2014), to address the challenges outlined above.

2 Methods

Here, we present FuMa, a computer program that reports identical fusion genes detected in RNA-seq, where matching is based on a user provided gene-name annotation. For two or more datasets, FuMa enlists all possible combinations of datasets that can be compared with each other. The iterative procedure starts by comparing all 2-dataset combinations. Every such comparison results in a new virtual dataset, containing only the matching fusion genes. Consequently, for comparisons with a larger number of datasets, input datasets and merged datasets will be compared with each other such that all possible combinations are tested, by comparing only two virtual datasets at a time.

Because several factors complicate matching using genomic positions, our solution is based on gene-name comparisons. Each breakpoint of a fusion gene consists of two genomic locations. We define the genomic locations of a breakpoint as *left* and *right*, where *left* < *right*, while sorting is applied first on chromosome name and, in case of equality, on genomic position. When the left and right locations are identical to the lexicographical order, we denote the acceptor–donor order as *forward*, otherwise as *reverse*. For each fusion gene a list of genes overlapping each associated genomic location from the user provided gene annotation (BED format) is added with the HTSeq library (Anders *et al.*, 2015). This step ensures that all fusion genes are annotated with consistent genomic identifiers rather than those provided by the detection tools themselves. Since genes frequently overlap and multiple genes may be annotated upon one location, we add genes as a list to use them for set-theory-based matching rather than exact gene matching (EGM). FuMa has two matching methods, *subset*-based matching (FuMa-s, default) and *overlap*-based matching (FuMa-o) further explained in Supplementary Section S1 and S2. Using FuMa-s, matching within any two datasets (both input- versus input- and input- versus merged dataset) is applied as follows:

- For each fusion gene in both datasets, remove entries that do not have gene annotations associated to both locations.
- Per dataset, merge duplicates such that a dataset contains only unique fusion genes. Two fusion genes are considered a duplicate by the *match*-function, which will later be explained as criterion used for matching fusion genes.
- Iterate over all fusion genes in both datasets such that all fusion genes of the first dataset are compared with all fusion genes of the second dataset. Assessing whether two fusion genes are identical is done by the *match*-function where two fusion genes are considered identical if: *one of the left gene lists is a subset of the other left gene list AND one of the right gene lists is a subset of the other* (Supplementary Table S1). Depending on the chosen parameters, the order of the genomic locations (forward or reverse) or the strands of the breakpoint may be taken into account as additional constraints. The comparison of any two virtual datasets will produce a 'merged' dataset that only includes matched fusion genes present in both input datasets. When two fusion genes match, the left and right gene sets of a matched fusion gene will be the intersect of the left or right gene sets of the input fusion genes. The intersect is chosen over the union to prevent gene lists from 'growing' after multiple iterations of matching (Supplementary Section S2).

3 Results

FuMa was tested on publicly available data (Berger *et al.*, 2010; Edgren *et al.*, 2011) and results provided as part of the Chimera package (Beccuti *et al.*, 2014) (Supplementary Section S6). Concordance of the matching methods was assessed using RefSeq gene annotation. While EPM reported less than half of the overlaps of FuMa, EGM performed better but was still outmatched by FuMa. Specifically, EGM missed 11.1–15.4% of the fusion genes due to not accounting for overlapping genes (Fig. 1) and importantly, five of the missed fusions had been validated. FuMa-o also reported a matching intergenic fusion event in a large gene that likely represents a false positive.

4 Discussion and conclusion

Accurate comparisons of identical fusion genes between different algorithms are desirable to increase confidence in *in silico* predictions and to allow performance analyses as well as in-depth evaluations of the algorithms used. Therefore, we developed FuMa, a software package that makes use of a gene name and set-theory-based strategy, taking into account the transcriptome to reduce uncertainty and produce a human and computer understandable output (Supplementary Section S5). FuMa is publicly available, available for Galaxy (Goecks *et al.*, 2010) and available as R package compatible with Chimera (Beccuti *et al.*, 2014). FuMa focuses on comparing breakpoints within annotated genes and is more sensitive compared with EPM and EGM. In addition, FuMa can handle intermediate results of several detection tools and thereby allows an evaluation of the interim steps of an algorithm. Last, we find limited overlap between ChimeraScan, Defuse and FusionMap (Ge *et al.*, 2011; Iyer *et al.*, 2011; McPherson *et al.*, 2011) in the Edgren dataset (Supplementary Fig. S4), which is in line with earlier reports (Beccuti *et al.*, 2014) indicating that further improvements in detecting fusion genes in RNA-seq data are needed.

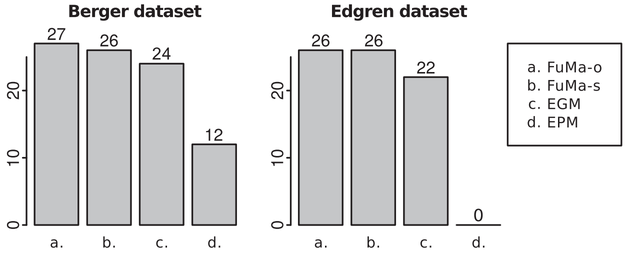

Fig. 1. Differences between the matching approaches in the Berger (left) and Edgren (right) dataset. Each bar represents the number of fusion genes found in two or more samples (Supplementary Section S6). For this analysis a RefSeq gene annotation was used

Funding

This study was performed within the framework of the Center for Translational Molecular Medicine (CTMM), TraIT project (grant 05T-401).

Conflict of interest: none declared.

References

Anders,S. *et al*. (2015) Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Beccuti,M. *et al*. (2014) Chimera: a bioconductor package for secondary analysis of fusion products. *Bioinformatics*, **30**, 3556–3557.

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

Carrara,M. *et al*. (2013) State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC Bioinformatics*, **14**(Suppl 7), S2.

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

Ewing,A.D. *et al*. (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.

Ge,H. *et al*. (2011) Fusionmap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, **27**, 1922–1928.

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*., **11**, R86.

Goode,D. *et al*. (2013) A simple consensus approach improves somatic mutation prediction accuracy. *Genome Med*., **5**, 90.

Iyer,M.K. *et al*. (2011) Chimerascan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903–2904.

Layer,R.M. *et al*. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*., **15**, R84.

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

Sanna,C.R. *et al*. (2008) Overlapping genes in the human and mouse genomes. *BMC Genomics*, **9**, 169.

Zhao,S. and Zhang,B. (2015) A comprehensive evaluation of ensembl, refseq, and ucsc annotations in the context of rna-seq read mapping and gene quantification. *BMC Genomics*, **16**, 97.