Advancing fusion gene detection from RNA-seq data

Konstantin Okonechnikov, Thomas F. Meyer, Fernando García-Alcalde

Max Planck Institute For Infection Biology, Berlin, Germany okonechnikov@mpiib-berlin.mpg.de

Gene fusion events and other types of chromosomal translocations are known to be related to several types of cancers and can be interpreted as hallmarks for the corresponding development stage of the disease. Therefore the detection of these translocations is attracting interest in modern genomics [1]. RNA-seq technology has achieved an unprecedented throughput and resolution which makes it one of the most promising approaches to detect gene fusions on a genome-wide scale. Unfortunately, RNA-seq data analysis can be challenging due to the complexity and amount of information to be processed, the error rate of the technology and the repetitive elements present in the genomes. Some computational methods have been proposed to deal with these challenges and proved to be useful in detecting gene fusions [3,4]. However, recent analysis demonstrated inconsistency of results between commonly used methods applied to the same validated datasets [2]. Thus, there is room for further improvement. Here we present an exploratory study on the problem of fusion detection and propose a novel approach, which allows a more efficient detection of transcriptome rearrangements, including novel fusions.

The typical input for fusion detection algorithms are the raw sequencing reads or their spliced alignments, together with a reference transcriptome annotation. There exist two main approaches to find initial candidates of possible fusion events. Methods such as deFuse[4] start with discovery of alignments of paired reads with discordant insert size to detect possible translocations. After the discordant pairs are found, initial fusion candidates are formed and directed spliced alignment of unmapped reads is applied to find the exact breakpoint of a fusion. Such methods can work only with paired-end data and rely heavily on the gene annotations. Other tools, like FusionMap[5] use spliced alignment to detect reads that span junction between genomic regions from different chromosomes. Using these split alignments the initial fusions are formed. Because of the repetitive nature of the genome and small size of the short reads this method may result in a large number of false positives. Our proposal combines both techniques to find initial list of fusion candidates. First, reads are mapped to the transcriptome using a regular short read aligner or to genome using splice-aware aligner. After that local alignment of unmapped reads to genome sequence is performed to find split evidence for the fusion events. If the data is paired-end, the initial alignment is analyzed to find discordant alignments to find supporting spanning read pairs. The candidates from both analyses are then clustered in an efficient manner to detect possible fusion events. Similarly to other existing methods, a number of filters are applied to detected putative fusions in order to increase the specificity of the results. To further improve the specificity of predictions the biological relevance of possible fusion event is estimated by taking into account the expression pattern of the involved genes and applying a probabilistic model for multi-mapped reads.

We have started to implement the described approach and tested its current implementation on in-house simulated RNA-seq datasets with fusion events and also on simulated datasets from FusionMap paper. In all cases our method demonstrated prediction accuracy equal or higher in comparison to other methods. We also performed tests on existing datasets from prostate and lung cancers and were able to detect with high specificity all fusions, that were confirmed earlier by experiments.

1. F. Mitelman, F. B. Johansson, F. Mertens (2007) "The impact of translocations and gene fusions on cancer causation." *Nature Reviews Cancer* **7.4:** 233-245.

2) M. Carrara et al. (2013) "State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity." *BioMed Research International 2013*

3) C.A. Maher et al. (2009) "Chimeric transcript discovery by paired-end transcriptome sequencing." Proceedings of the National Academy of Sciences **106.30:** 12353-12358.

4) A. McPherson et al. (2011) "deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data." *PLoS computational biology* **7.5:** e1001138.

5) H. Ge , et al. (2011) "FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution." *Bioinformatics* **27.14**: 1922-1928.