EricScript to identify gene fusions in paired-end RNA-seq data

Benelli M(1,2), Pescucci C(1), Marseglia G(1), Severgnini M(3), Torricelli F(1), Magi A(2,4)

(1) Diagnostic Genetic Unit, Careggi University Hospital, Firenze, Italy
(2) Center for the Study of Complex Dynamics, University of Firenze, Italy
(3) Institute of Biomedical Technologies (ITB), National Research Council (CNR), Segrate, Italy
(4) Laboratory of Bioinformatics, Department of Medical and Surgical Critical Care, University of Firenze, Italy

Contact: matteo.benelli@gmail.com

Motivation

The identification of genomic rearrangements in cancer research plays a main role to investigate causes and development of the disease. Gene fusions are common alterations in which two genes are fused, leading to the production of a chimeric transcript that may have a new or altered activity. Gene fusions are well-known mechanisms for oncogene activation in leukemias, lymphomas and sarcomas but recent studies found novel chimeric transcripts also in common epithelial cancers such as prostate cancers and non-small-cell lung cancer. The last few years have seen the emergence of several high-throughput sequencing (HTS) platforms that enable to sequence hundreds of millions of short sequences (reads) simultaneously and have routinely been applied to genome, epigenome and transcriptome studies.

At present, several computational approaches have been developed for the detection of chimeric transcripts in RNA-seq data. Although the most of them shows good sensitivity in discovering chimeric transcript events, much work remains to reduce the huge number of false positives that arises from this kind of analysis. Moreover, only very few methods are able to detect gene fusions involving non-canonical splice junctions.

Methods

Here we present a computational method, named EricScript (chimERIC tranSCRIPT detection algorithm), for the detection of gene fusions in PE RNA-seq data. A easy-to-use command line interface enables user to detect gene fusions for a variety of species (i.e., all the species available in ensembl database), to generate synthetic gene fusions and estimate detection and scoring accuracy (see Results section for more details). EricScript is a computational framework that uses a combination of four alignment processes to identify fusion transcript signatures. It comprises the following steps:

1. Mapping of the reads against the transcriptome.
2. Identification of discordant alignments and building of the exon junction reference.
3. Recalibration of the exon junction reference.
4. Scoring and filtering the candidate gene fusions.

The first alignment (performed by BWA (Li et al, 2009) is used to identify discordant alignments and to build a exon junction reference. This step is followed by the mapping of all the reads against this novel reference to detect reads that are not properly mapped (i.e., partially mapped and unmapped reads). In order to precisely estimate the exact borders of the junctions, our method performs a further local realignment (performed by BLAT (Kent 2002)) of the not properly mapped reads against the exon junction reference. The last mapping procedure allows for the identification of the spanning reads, that is the reads that span across the junctions, and produce a list of candidate fusions. Finally, EricScript estimates a probability score (adaboost classifier) for each predicted fusion and uses several heuristic filters to remove analysis artefacts.

Results

To assess a reliable estimation of the performance of EricScript, we simulated paired end (PE) RNA-seq data with synthetic gene fusions and we compared our method with ChimeraScan (Iyer et al., 2011), DeFuse (Mcpherson et al., 2011), FusionMap (Ge et al., 2011) and ShortFuse (Kinsella et al,
15,000 Synthetic gene fusions were generated by randomly sampling 5'-transcripts and 50 3'-transcripts from the Ensembl Transcriptome database version 65. Each sampled 5'-transcript was joined with the corresponding 3'-transcript and the breakpoints for both transcripts were randomly chosen among all the known splicing sites of synthetically fused genes (Intact Exons) or without exploiting information of the known splicing sites of synthetically fused genes (Broken Exons). Then, we simulated 50, 75 and 100bp PE reads by means of wgsim (http://github.com/lh3/wgsim) and we varied the number of reads generated by wgsim in order to simulate different levels of coverage (from 1 to 50). We compared the performance of the gene fusion algorithms on synthetic data by several statistical indices that allowed us to study detection sensitivity and specificity and scoring accuracy (the ability of such an algorithm in discriminating between true and false positive events). The results we obtained from the synthetic study has been published in a recent paper (Benelli et al., 2012) and showed that EricScript performs generally better than compared algorithms. We also applied our algorithm to five publicly available datasets and we tested its capability in rediscovering previously characterized gene fusions. Our analyses on both synthetic and real data demonstrated that EricScript is very reliable in assembling the correct sequence of fusion junctions, allowing for the detection of chimeric events with a resolution of 1bp.