Bioinformatics Advance Access published October 23, 2012

Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript.

Matteo Benelli [1,2]*, Chiara Pescucci [2], Giuseppina Marseglia [2], Marco Severgnini [3], Francesca Torricelli[1,2] and Alberto Magi[2,4]

[1] Diagnostic Genetic Unit, Careggi University Hospital, 50134 Florence, Italy, [2] Center for the Study of Complex Dynamics, University of Florence, 50019 Florence, Italy, [3] Institute of Biomedical Technologies of the National Research Council (ITB-CNR), 20090 Segrate, Italy, [4] Department of Medical and Surgical Critical Care, University of Florence, 50134 Florence, Italy.

Associate Editor: Dr. Matteo Benelli

ABSTRACT

Motivation: The discovery of novel gene fusions can lead to a better comprehension of cancer progression and development. The emergence of deep sequencing of trancriptome, known as RNA-seq, has opened many opportunities for the identification of this class of genomic alterations, leading to the discovery of novel chimeric transcripts in melanomas, breast cancers and lymphomas. Nowadays, few computational approaches have been developed for the detection of chimeric transcripts. Although all of these computational methods show good sensitivity, much work remains to reduce the huge number of false positive calls that arises from this analysis.

Results: We proposed a novel computational framework, named chimEric tranScript detection algorithm (EricScript), for the identification of gene fusion products in paired-end RNA-seq data. Our simulation study on synthetic data demonstrates that EricScript enables to achieve higher sensitivity and specificity than existing methods with noticeably lower running times. We also applied our method to publicly available RNA-seq tumour datasets and we showed its capability in rediscovering known gene fusions.

Availability: The EricScript package is freely available under GPL v3 license at `http://ericscript.sourceforge.net`.

Contact: matteo.benelli@gmail.com

1 INTRODUCTION

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 (Mitelman *et al.*, 2007) but recent studies found novel chimeric transcripts also in common epithelial cancers such as prostate cancers (Tomlins *et al.*, 2005) and non-small-cell lung cancer (Soda *et al.*, 2007). 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 being applied to genome, epigenome and transcriptome studies.

The sequencing of trancriptome (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008), known as RNA-seq, has been widely used for the study of abundance estimation (Jiang and Wong, 2009), RNA editing (Picardi *et al.*, 2010), identification of novel transcripts (Robertson *et al.*, 2010) and splicing variants detection (Trapnell *et al.*, 2010; Wang *et al.*, 2010). In 2009, Maher and collegues (Maher *et al.*, 2009) proposed a new methodology to comprehensively catalog functional gene fusions in cancer by using paired-end (PE) transcriptome sequencing data. The basic idea behind their approach was the identification of paired reads mapping against different genes (discordant alignments). Following this original approach, several studies for the detection of gene fusions have been carried out: Pflueger *et al.* (2011) found non-ETS gene fusions in human prostate cancer, Berger *et al.* (2010) identified 11 novel melanoma gene fusions produced by underlying genomic rearrangements and 12 novel read-through transcripts, Edgren *et al.* (2011) detected 24 novel and 3 previously known fusion genes in breast cancer cells and Steidl *et al.* (2011) found highly expressed gene fusion involving the major histocompatibility complex class II transactivator CIITA in KM-H2 cells.

At present, several computational approaches have been developed for the detection of chimeric transcripts in RNA-seq data. FusionSeq (Sboner *et al.*, 2010) identifies gene fusions by means of a two step analysis: identification of potential fusions based on PE mapping and the application of a sophisticated filtration cascade to filter out analysis artifacts. DeFuse (McPherson *et al.*, 2011) is a software package that uses clusters of discordant paired end alignments to perform a split read alignment analysis for finding fusion boundaries. ChimeraScan (Iyer *et al.*, 2011) is a tool that implements the original computational methodolgy followed by Maher *et al.* (2009). FusionMap is able to search for gene fusion products in both single-end and paired-end sequencing by using "seed reads" (Ge *et al.*, 2011). TopHat-fusion implements several changes to the TopHat aligner, all designed to enable the discovery of fusion transcripts (Kim and Salzberg, 2011). ShortFuse (Kinsella *et al.*, 2011) detects chimeric transcripts RNA-seq data by using both unique and ambiguously mapping read pairs. Although all of these computational methods show good sensitivity in discoverying chimeric transcript events, much work remains to reduce the huge number of false positives that arises from this kind of analysis.

Here we present a novel computational method, named EricScript (chimERIC tranSCRIPT detection algorithm), for the detection

*to whom correspondence should be addressed

of chimeric transcripts in PE RNA-seq data. The novelty of our approach consists in an efficient recalibration process of the exon junction reference that enables to increase sensitivity and specificity and to reduce running times. Moreover, we introduce a set of scores that enable to distinguish with high precision between true chimeric transcripts and false positive events and reduce the large amount of calls generated from data analyses. To evaluate the performance of EricScript, we generated synthetic datasets of different read length simulating different levels of coverage and we compared it with other four state-of-the-art algorithms. The synthetic datasets were also used to train an adaptive boosting (AdaBoost) classifier to rank analysis results. We implemented the procedure that we used to simulate gene fusions in the EricScript package. Our simulation study demonstrates that EricScript enables to achieve higher sensitivity and specificity than existing methods with noticeably lower running times. We also applied our method to publicly available PE RNA-seq tumour dataset and we showed its capability in rediscovering known gene fusions.

2 METHODS

EricScript is a computational framework (see Figure 1) that uses a combination of four alignment processes to identify fusion transcript signatures. It comprises the following steps:

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

The first alignment (performed by BWA (Li and Durbin, 2009)) is used to identify discordant alignments and to build an 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 realigment (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 for each predicted fusion and uses several heuristic filters to remove analysis artifacts.

2.1 Transcriptome reference

EricScript includes a pre-built transcriptome reference for the alignment process and the retrieval of information about genes. The set of genes we consider for the analysis of chimeric transcripts is created by including the Ensembl genes with the HUGO Gene Nomenclature Committee (HGNC) identifier (Seal *et al.*, 2011). The sequences of each gene are built by joining the sequences of the exons from the different transcript isoforms (exon-union model). We distribute EricScript with the latest version of Ensembl Genes (`http://www.ensembl.org`) and, therefore, no other reference is required to perform the analysis.

2.2 Identification of discordant reads

The first essential step in chimeric transcripts identification from PE RNA-seq data is to select reads for which each mate of the fragment aligns against different genes with opposite orientation. Although our approach is independent by the short read aligner, we decided to use the Burrows-Wheeler Alignment (BWA) tool since it reaches the best balance between sensitivity, specificity and computational time (Ruffalo *et al.*, 2011). The BWA mapping of all the reads against the pre-built Ensembl transcriptome is used to identify discordant alignments. In order to increase the sensitivity of BWA in discovering discordant alignments (especially when the length of the reads is greater equal than 75nt), the parameter $ntrim$ allows Ericscript to trim PE reads to a selected value only for this alignment (see Supplemental Material). EricScript enables to choose discordant alignments supported by a minimum number of the reads ($minreads$) and a minimum mapping quality of the supporting reads ($MAPQ$). The set of reads that map to the same pairs of genes in the same orientation are considered by EricScript to build a reference of putative gene fusion events (see Figure 2a and Supplemental Material). Discordant alignments between paralogous genes (Ensembl Paralogous Human Genes) are filtered out and not considered for the downstream analysis.

The usage of exon-union model together with BWA allows us to exploit the MAPQ's BWA estimation. In fact, the quality of the mapping assigned by BWA to each alignment is an excellent way to reduce false positives of the mapping process (Ruffalo *et al.*, 2011). Moreover, exon-union model reports all the exons once allowing us to completely exploit the MAPQ information, mainly by excluding discordant alignments with MAPQ lower than a selectable threshold value.

2.3 Candidate exon junction reference

The identification of discordant alignments enables to build the candidate exon junction reference that will be used to search for split-read signatures. For each set of discordant alignments (i.e, all the discordant reads that map against the same couple of genes), 5'-gene is identified by the signature of forward reads (first reads of the pairs) mapping to it, while the reverse reads are exploited to identify 3'-gene. Let R_5 (R_3) be the genomic region encompassing all the alignments of crossing reads to 5'-gene (3'-gene). Let the m (n) candidate exons for 5'-gene (3'-gene) be all exons overlapping R_5 (R_3). As schematically illustrated in Figures 2b and 2c exon junction reference is generated by joining together the sequences of the m exons of 5'-gene with the n exons of 3'-gene, according to the strand of transcription of both the genes. Bearing this in mind, the candidate exon junction will result from the union of the last nucleotide of the m-th exon with the first

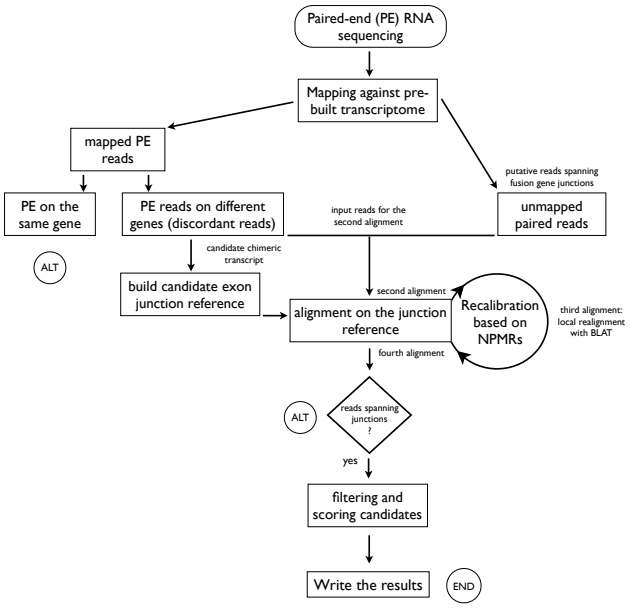

Fig. 1. The computational pipeline of EricScript.

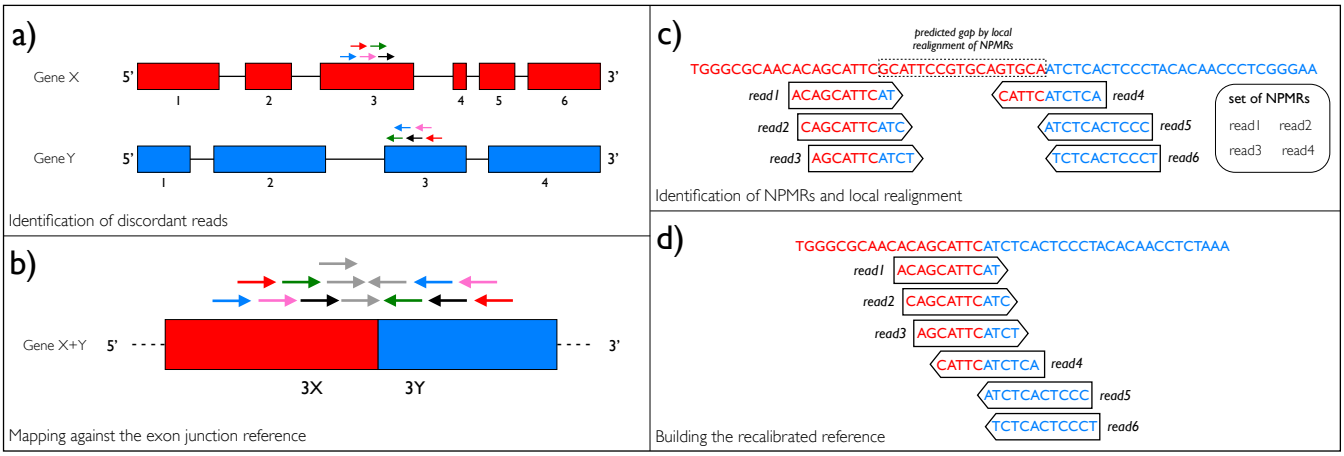

Fig. 2. A semplified scheme for illustrating the recalibration procedure of EricScript. a) Identification of discordant alignments and construction of the exon junction reference. The reads with the same color belong to the same cDNA fragment, that is they are mates. b) Mapping of all the short reads against the exon junction reference by means of the BWA aligner. The not properly mapped reads (NPMRs) are reported in grey color. c) Identification of NPMRs and local realignment of them against the exon junction reference by using BLAT. The local realignment reports a gap d) that allows EricScript to recalibrate the exon junction reference. The mapping of all the reads against the augmented recalibrated reference (see text) enables the identification of the junction spanning reads.

nucleotide of the n-th exon, according to the strand of transcription of both the genes.

2.4 Recalibration of the exon junction reference

Gene fusions can involve both fusions between genomic boundary of exons and fusions between any genomic position of the two exons. Moreover, since we build the set of putative fusions by joining the exon boundaries of the candidate 5'-3' fused genes, our junction reference strictly depends on the Ensembl transcriptome. Bearing this in mind, all the reads are mapped against the exon junction reference and the reads that are not properly mapped (not properly mapped reads, NPMRs) are identified. We define NPMRs as the reads that map for a fraction of their length against the reference or are unmapped (see Figure 2b). NPMRs represent the candidate set of reads that span a fusion boundary and allow us for finding fusions that involve middle of exons. Our pipeline classifies reads as NPMRs if they are unmapped or their string for mismatching positions (that is, MD:Z tag of SAM file) reports a mismatch and its mapping quality is greater than 0. Each NPMR is then locally re-aligned against its corresponding junction by means of the BLAT aligner in order to predict the existence of gaps greater than 3 bp (see Supplemental Material for more details). As schematically illustrated in Figure 2b and Figure 2c, this step allows us to investigate if the majority of the NPMRs predicts a gap. This means that when two or more gene fusion isoforms are expressed in a sample, EricScript is able to only detect the trascript with the highest espression level. The exon junction reference is then recalibrated by taking into account the predicted gap (see Figure 2d).

2.5 Mapping against the recalibrated reference

After the recalibration step we build a novel reference that comprises both the pre-built Ensembl Transcriptome and the previously recalibrated junction reference. All the reads are mapped against this augmented reference by means of the BWA aligner. Putative gene fusion junction are selected for downstream analysis if there exists at least one read that spans the junction.

2.6 Scoring the candidate fusions

As already reported by Edgren *et al.* (2011), we expect genuine fusion junctions to be characterized by a ladder-like pattern of short reads alignment across the junctions. On the other hand, a typical false positive event due to a "wrong" pattern is represented by short reads aligning around to the same position (shifted at maximum of 2-3 bp). In order to distinguish genuine from wrong patterns we introduced three novel scores, named genuine junction score (GJS), Edge Score (ES) and Uniformity Score (US). In order to produce a single score we trained an AdaBoost classifier on the aforementioned scores that allows EricScript to rank predicted gene fusions. To comprehend the parameters used to define these score, refer to Figure 3.

Genuine Junction Score. The aim of introducing GJS is to assign higher scores to those gene fusions characterized by the presence of reads spanning a comparable number of bases in both the genes. For each junction j, duplicated spanning reads are considered only once. This set of $n_{j,\text{unique}}$ reads is used to calculate the GJS in the following:

$$GJS_j = \frac{\sum_{i=1}^{n_{j,\text{unique}}} N(x_i|\mu,\sigma)}{\sum_{i=1}^{n_{j,\text{unique}}} N(z_i|\mu,\sigma)}, \quad (1)$$

where x_i is the relative position of each read with respect to the junction, N is the normal distribution with mean $\mu = -rl/2$ and standard deviation $\sigma = rl/4$ (rl is the readlength of the reads) and z is a vector made up of $n_{j,unique}$ "sham" positions that would have minimum distance from $-rl/2$, that is the position for which reads span both the genes with the maximum number of bases. The choice of using z_i enables to constrain GJS_j between 0 and 1. To be clear, we report the following example: let assume to have two cases: a) $rl = 50$, $x = (-1, -5, -40, -5)$

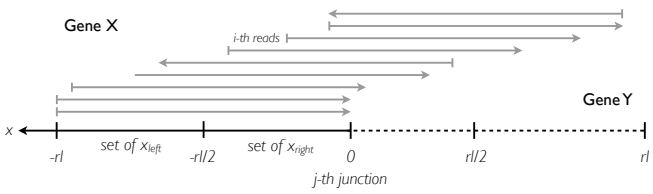

Fig. 3. Coordinate system and variables for evaluating the scores of EricScript.

and b) $rl = 50$, $x = (-10, -25, -30)$. For both the cases, the vector $z = (-25, -24, -26)$. z has 3 elements as the number of unique positions for both the examples we reported ($n_{j,unique} = 3$) and its elements are the positions with minimum distance for $-rl/2$. In case a) we obtain a GJS = 0.31 while in case b) GJS = 0.81.

Edge Score. For each junction j, we parted reads in two sets: reads with relative position $x > -rl/2$ with respect to the junction (we marked them by the $right$ subscript) and reads with $x \leq -rl/2$ (we marked them by the $left$ subscript). We defined ES by the following formula:

$$ES_j = 1 - 1.1^{\left(\frac{-\overline{x_{j,\text{left}}+rl}+\overline{x_{j,\text{right}}}}{2}\right)}, \quad (2)$$

where the overlines represent the mean and the 1.1 base is arbitrarily chosen to soften variations in the exponent. The ES score allows us to give lower score values to events with the majority of reads that fall in proximity of the fusion junction or in proximity of $-rl$.

Uniformity Score. The US score was conceived to assign higher scores to events in which the number of spanning and crossing reads are comparable. For each junction j, US is defined as the following expression:

$$US_j = \frac{\min(n_{j,cross}, n_{j,span})}{\max(n_{j,cross}, n_{j,span})}, \quad (3)$$

where $n_{j,cross}$ and $n_{j,span}$ are the number of crossing and spanning reads, respectively.

AdaBoost classifier. In order to better rank predicted gene fusions, we used an AdaBoost classifier trained with synthetic data (see Results Section and Supplemental Material). As already reported by McPherson *et al.* (2011) AdaBoost was selected because enables to improve the predictive power of each individual score and summarize the aforementioned measures into a single score (we indicate it as "EricScore").

2.7 Filtering the results

Filtering is an essential procedure when dealing with the detection of chimeric transcripts, since several types of noise of both sequencing and analysis process can lead to the detection of a large amount of gene fusion artifacts (Sboner *et al.*, 2010; Edgren *et al.*, 2011). To this end, we designed a set of heuristic filters with the aim of discarding these false positive events.

Duplicate reads. We discard all the PE reads that exactly map to the same position since they may derive from PCR or optical artifacts. We use the command *rmdup* of samtools to remove these events.

Pattern of short reads. Scoring the candidate fusions by means of $EricScore$ allows us to assign to each candidate a probability score of "well" pattern and thus classify all the fusions for discriminating between real transcripts and false positive events.

Transcript similarity. Reads mapping on homologue regions of different genes can lead to chimeric transcript artifacts. To minimize these events, we use BLAT to map the 100 bp sequence region around the wild type junction against the Ensembl transcriptome. If BLAT finds that the 100 bp window sequence map $\geq 80\%$ of its length against one of the two candidate fused genes, we remove the candidate fusion.

Junction homology. The junction coming from fusion process can be homologue with other regions of the transcriptome. To take into consideration these events, we map the 100 bp sequence region around the predicted junction against the Ensembl transcriptome with BLAT. If BLAT finds a homology with other genes, EricScript reports the percentage of homology, that is the percentage of the bases of the homologue gene(s) that overlaps the 100 bp junction sequence.

2.8 Writing results

After the filtering process, EricScript reports the candidate fusions in two tab-delimited files: one file contains all the predicted fusions while the other reports the fusions with $EricScore > 0.5$. For each predicted gene fusion, EricScript outputs several information that include the names of 5' and 3' genes and their corresponding biological descriptions, the breakpoint positions for both the genes, the sequence that arises from the fusion process and the type of fusion (inter-chromosomal, intra-chromosomal, read-through or CIS-acting transcripts). Moreover, we report the 4 scores (GJS, ES, US, $EricScore$) and the estimation of gene expression of wild type genes and of the gene fusion product by using a read count approach (see Supplemental Material for more details).

2.9 Implementation, requirements and availability

EricScript is written in perl, R and bash scripts. It requires the BWA aligner to perform the mapping of the PE RNA-seq short reads against

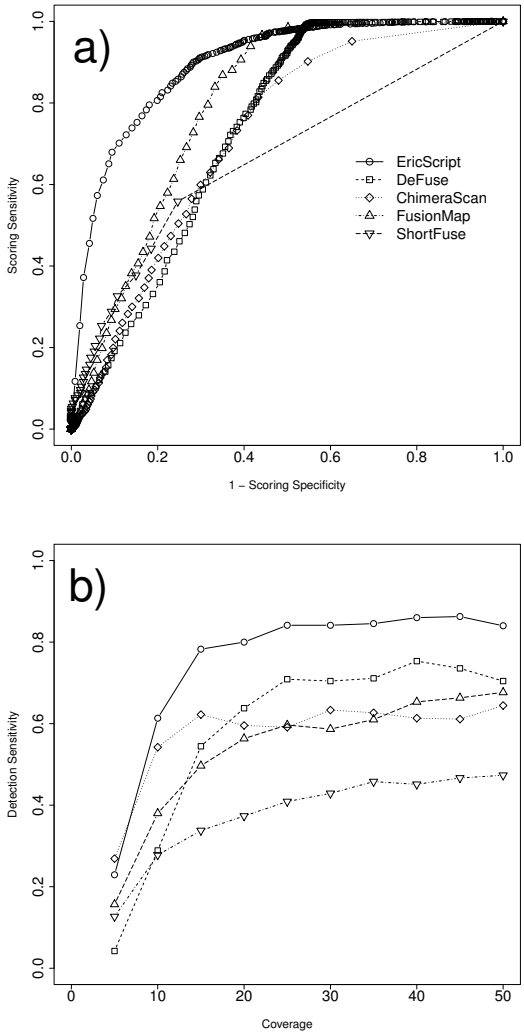

Fig. 4. Results of the simulation study among 150 synthetic IE datasets. a) Comparison between ROC curves obtained for EricScript and the other state-of-the-art gene fusion detection methods. b) TPR of EricScript and the other fusion discovery methods versus coverage (each point of the plot represents a bin of 5 values of coverage). The legend of a) is relative to both the plots.

the transcriptome, the SAMtools software package (Li *et al.*, 2009) to handle with the SAM/BAM files created during the analysis and the BLAT tool to perform the local realignment of the NPMRs against the exon junction reference. EricScript is freely available under GPL v3 license at http://ericscript.sourceforge.net.

3 RESULTS

Synthetic data

To assess a reliable estimation of the performance of EricScript, we simulated PE RNA-seq data with syntethic 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.*, 2011). We generated each syntethic dataset with the following recipe: we randomly extracted two millions of short reads from the RNA-seq data of untreated human pulmonary microvascular endothelial cells generated by Zhang *et al.* (2012) (SRA accession code: SRX099065). This dataset is made of 10.3G PE 100bp reads sequenced by the Illumina HiScan SQ (Illumina Inc., San Diego, CA 92122 USA). By aligning all the reads against the Ensembl Transcriptome database version 65 with BWA (version 0.6.2-r126), we estimated that cDNA fragments were generated from cDNA fragments of length $\sim$ 164 and standard deviation $\sim$ 48. The reads were also trimmed to 50bp and 75 bp to evaluate the performance of each algorithm for different read lengths. The purpose of introducing these reads in our study is to simulate a background of "synthetic" gene activity. To simulate synthetic gene fusion products, we sampled 50 5'-transcripts and 50 3'-transcripts from the Ensembl Transcriptome database version 65. We created two distinct datasets:

- Intact exons (IE): 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.
- Broken exons (BE): Each sampled 5'-transcript was joined with the corresponding 3'-transcript and the breakpoints for both transcripts were randomly chosen without exploiting information of the known splicing sites of synthetically fused genes.

From these novel references, we simulated 50, 75 and 100bp PE reads by means of wgsim (`http://github.com/lh3/wgsim`) (with -d 164 -r 0.0001 -R 0.001 -s 48). We varied the number of reads generated by wgsim in order to simulate different levels of coverage (from 1 to 50). The final synthetic PE RNA-seq dataset is built by merging, for each read length data, the background dataset and the simulated gene fusions (for both IE and BE data). Although such a synthetic dataset is an ideal and simplistic case for simulating gene fusion processes, the use of it allows us to objectively assess chimeric trancripts discovery algorithms. To this end, we generated 50 synthetic PE RNA-seq datasets for each read length data and for both BE and IE events (for a total of 300 synthetic datasets and 15000 synthetic gene fusions) and we analysed them by using EricScript (with $minreads = 2$, $MAPQ = 1$ and $ntrim = 50$), ChimeraScan, DeFuse, FusionMap and ShortFuse (see Supplemental Material and Supplemental Table 1 for more details). We compared the performance of these algorithms by using the following statistical indices:

Table 1. Comparison of statistical indices between EricScript and the other gene fusion detection methods among the 150 synthetic IE datasets. All the values are obtained by averaging across all the simulations we performed.

Method	TPR	FPR	FNR	TPSR[1]	AUC	Time[2]
EricScript	0.75	0.15	0.25	> 0.99	0.90	0.53
ChimeraScan	0.58	0.40	0.42	-	0.72	0.76
DeFuse	0.58	0.39	0.42	> 0.99	0.76	1.3
FusionMap	0.54	0.63	0.46	> 0.99	0.80	1.6
ShortFuse	0.38	0.13	0.62	-	0.67	0.33

[1] ChimeraScan and ShortFuse do not output fusion junction sequence.
[2] Expressed as CPU hours.

- True Positive Rate (TPR) or *detection sensitivity*. We defined TPR as the number of gene fusions correctly predicted by the algorithm divided by the total number of simulated fusions (50).
- False Positive Rate (FPR) or *detection specificity*. We defined FPR as the number of predicted gene fusions that are not in the list of simulated fusions divided by the total number of detected events.
- False Negative Rate (FNR). FNR corresponds to the number of undetected gene fusions divided by the total number of simulated fusions (that is, 1- *detection sensitivity*).
- True Positive Sequence Rate (TPSR). TPSR is the number of correctly determined junction sequences divided by the number of correctly predicted gene fusions.
- Area Under the Curve (AUC). AUC is a measure of the accuracy of each algorithms in discriminating between true and false positives. This parameter is estimated by means of the Receiver Operating Characteristic (ROC) curve. Details on how ROC curves were built are available in Supplemental Material.

The TPR, FPR, FNR statistical indices are useful to estimate "detection accuracy" of each algorithm. In fact, these measures considered all the calls, irrespective of scores assigned to the identified fusion events. On the other hand, AUC and the ROC curves reported in this manuscript enable to evaluate "scoring accuracy" of each algorithm, that means the ability of such an algorithm in discrimating between true and false positive events.
The results of these analyses are reported in Figure 4, Table 1 and Supplemental Material. The ROC curves of Figure 4a and Supplemental Figure 2 clearly show that our algorithm obtains better performance than the other state-of-the-art methods in distinguishing between true and false positive events. EricScript outperforms the other methods in all the simulations we performed with different read lengths for both BE and IE datasets, with the exception of data with read length equal to 75bp and 100bp in which FusionMap reaches similar results. Figure 4b shows the capability of the five algorithms in identifying the correct fusion genes versus the simulated coverage for all the IE datasets, while Supplemental Figure 2b is related to BE simulations. For both the datasets, when coverage is smaller than 10x, all the algorithms

are not able to reliably discover fused transcripts. On the other hand, for IE data and for coverage greater than 10x, our method detects gene fusions with TPR higher than 0.8 followed by DeFuse that is able to discover almost 70% of the fusions we simulated. ChimeraScan and FusionMap detect around 60% of gene fusion events. For BE data (see Supplemental Figure 2b), ChimeraScan and ShortFuse lose their prediction capability while EricScript, DeFuse and FusionMap do not. The strong performance of EricScript and DeFuse in BE data is due to the fact that both the algorithms detect fused transcripts by a split-read approach allowing them to identify fusions involving middle of exons. FusionMap performs well on these datasets since we run it with $G = 0$ to not penalize non-canonical splice patterns (see Supplemental Material).

Conversely, ChimeraScan and ShortFuse are computationally designed to privilege fusions involving known splicing sites. The results corresponding to different read lengths are reported in Supplemental Figures 3-8. In these plots we observe that the overall performance of Defuse and ShortFuse decrease while the length of the reads increases. This is due to two main reasons: *i)* these algorithms have been calibrated on reads of 50 nt in size and *ii)* at fixed coverage, the longer are the reads and the smaller is the number of discordant reads. Increasing read length does not affect ChimeraScan, EricScript and FusionMap performance. The results reported in Table 1 and Supplemental Table 6 are obtained by averaging across the 150 synthetic IE datasets and 150 synthetic BE datasets, respectively. The FPRs reported in these Tables highlight that our algorithm outperforms the majority of the other methods: the probability of EricScript to make a wrong call is about 0.15 while other algorithms obtain FPR values that range between 0.15 (ShortFuse) and 0.63 (FusionMap). However, it is important to note that FPR may be misleading since a tool could predict thousands of very low scoring false positives. This would affect FPR, even though these events are easily discernable from true positives based on their low score (see the comments above on ROC curves analysis) or on the fact that these events are supported by a very few number of supporting reads. In fact, when we consider calls with predicted number of supporting reads > 5, all the algorithms (especially FusionMap) show a strong increase in specificity (see values in parentheses of Supplemental Tables 7-12 and Supplemental Figure 9) to the detriment of a small decrease in terms of sensitivity. The simulation study we performed also shows that EricScript, DeFuse and FusionMap obtain excellent results in reconstructing the correct fusion gene junction sequences while ChimeraScan and ShortFuse do not output this information (TPSR score). Table 1 also reports the average computational time taken by each algorithm to complete the analysis: ShortFuse obtains the best performance and requires about 80% less time than FusionMap. Although EricScript uses a four-step alignment pipeline, it takes only 0.53 hours per CPUs to perform the analysis. This is due to the fact that we use a transcriptome instead of a genome reference to map the reads: this feature does not allow EricScript to detect fusions involving unannotated transcripts. Supplemental Tables 7-12 reports the results relative to different read lengths and make clear the aforementioned coverage effect. If we set $ntrim = 0$, also EricScript is affected by coverage effect mainly for read length equal to 100bp and coverage smaller than 10x (see Supplemental Tables 2-3). This is the reason why we performed these analyses with $ntrim = 50$ for read length equal to 75 and 100 bp (see Supplemental Material for more details).

Application to previously reported gene fusions

We applied EricScript to publicly available PE RNA-seq datasets (see Table 2) with the aim of evaluating its capability in discovering previously characterized gene fusion products. We analysed the NCI-H660 prostate cell line dataset for the TMPRSS2-ERG and FOXP1-RYBP fusions (Sboner *et al.*, 2010; Pflueger *et al.*, 2011) (see Supplemental Material for a comparison between EricScript and DeFuse on these data) and we searched for the 23 validated gene fusions in the four breast cancer cell lines of Edgren *et al.* (2011) (SRA accession: SRP003186). We run EricScript on these datasets with two different sets of input parameters: setting a with $minreads = 2$ and $MAPQ = 1$ and setting b with $minreads = 3$ and $MAPQ = 20$. In both cases, EricScript took about 14 CPU hours to complete the whole analysis on about 60 million reads. The results of the analyses are reported in Table 3.

With parameter setting a, our method predicted 489 fusions in total (see Supplemental File 1). It was able to detect 22 of the 25 known fusions and for all of them EricScript is able to assemble the correct sequence of the junction. In the BT-474 library we predicted 9/10 validated gene fusion while we missed the fusion CPNE1-PI3. This fusion was filtered out by EricScript since BWA found three discordant alignments with $MAPQ = 0$ between CPNE1 and both PI3 and RBM12, potentially indicating a read-through between CPNE1 and RBM12. This situation also happens for the ANKHD1-PCDH1 fusion in the SK-BR-3 sample. BWA found six discordant alignments with $MAPQ = 0$ between ANKHD1 and both PCDH1 and ANKHD1-EIF4EBP3 (ENSG00000254996). EricScript identified all the other validated fusions of SK-BR-3 dataset, including the DHX35-ITCH fusion that, as reported by Kim and Salzberg (2011), neither DeFuse nor TopHat-fusion are able to detect. In the KPL-4 cell line we were able to detect all the known fusions, while in the MCF-7 sample our method was not able to rediscover the RPS6KB1-TMEM49 fusion since BWA found no discordant read.

With setting b, EricScript predicted 20 of the 25 validated fusions (see Supplemental File 2) and missed the events supported by less than 3 supporting reads (WDR67-ZNF704 and PPP1R12A-SEPT10). In this case our method predicted 193 fusions in total.

Table 3 also shows that all the predicted known fusions with the exception of DHX35-ITCH and TATDN1-GSDMB have a $EricScore > 0.5$ (11/22 present $EricScore \geq 0.90$). In particular, the low score of TATDN1-GSDMB is due to a low value of US ($US \sim 0.29$, see Supplemental Files 1-2): US was introduced in order to assign a higher score to candidate fusion genes in which the number of junction spanning single reads and paired-end reads connecting the genes are similar. Although this is a valid assumption for most fusion genes, it may not be true for fusion genes in which only a short stretch of the 5' (or 3') gene is present. Moreover this measure is dependent on library specific factors including the length of the cDNA fragments and lengths of the reads. Generating a specific dataset that simulates these features for training our classifier would be useful to improve the classification power of $EricScore$. Despite of that, these results indicate that $EricScore$ is very reliable for discriminating between true and false positive calls also in real data. If we consider only the predicted gene fusions with $EricScore > 0.5$, we are able to significantly reduce the number of our set of calls: indeed, we found

Table 2. RNA-seq datasets used for EricScript validation. The number of gene fusions identified by EricScript with $EricScore > 0.5$ is reported between parentheses.

Reference	Library	Number of reads	Read length	Time (CPU hours)[1]	Predicted fusions setting a	Predicted fusions setting b
Sboner *et al.* (2010)	NCIH660	6,512,688	51	1.2	31 (7)	12 (5)
Edgren *et al.* (2011)	BT-474	21,423,697	50	3.8	193 (53)	84 (43)
Edgren *et al.* (2011)	SK-BR-3	18,240,246	50	3.4	180 (35)	61 (22)
Edgren *et al.* (2011)	KPL-4	6,796,443	50	1.1	39 (8)	15 (4)
Edgren *et al.* (2011)	MCF-7	8,409,785	50	1.4	46 (9)	21 (10)

[1] The reported run time is for EricScript with setting a.

a total of 112 fusions for setting a and a total 84 fusions for setting b.

4 CONCLUSION

In this work, we discussed a novel computational approach to use discordant alignments of paired-end RNA-seq data to identify chimeric transcripts. Our method, named EricScript, makes use of the local realignment of the sequence reads that align across a gene fusion boundary to search for evidence of gene fusion events. We introduced three novel scores for classifying the "goodness" of the distribution of the reads that span the junctions. The results we obtained demonstrate that these approaches, joined with the application of a filtering step, perform better than existing methods in distinguishing between real fusions and false positive events, resulting in a smaller but robust set of calls. In fact, the analyses we performed on the synthetic gene fusion datasets showed that EricScript obtains very good results in terms of both specificity and sensitivity with low computational times. Moreover, our synthetic study demonstrated that split read based methods (EricScript and DeFuse) obtain better performance than the other algorithms and

Table 3. EricScript results in the publicly available PE RNA-seq datasets of Edgren *et al.* (2011) and Sboner *et al.* (2010). The scores are relative to EricScript with setting a.

Library	5' Gene	3' Gene	Crossing reads[1]	Spanning reads[2]	EricScript setting a	EricScript setting b	EricScript correct sequence	EricScore
NCIH660	TMPRSS2	ERG	18	15	✓	✓	✓	0.97
NCIH660	FOXP1	RYBP	12	6	✓	✓	✓	0.57
BT-474	ACACA	STAC2	56	80	✓	✓	✓	0.89
BT-474	VAPB	IKZF3	41	32	✓	✓	✓	0.97
BT-474	ZMYND8	CEP250	36	25	✓	✓	✓	0.96
BT-474	RAB22A	MYO9B	10	21	✓	✓	✓	0.94
BT-474	SKA2	MYO19	8	9	✓	✓	✓	0.97
BT-474	STARD3	DOK5	6	5	✓	✓	✓	0.93
BT-474	LAMP1	MCF2L	5	2	✓	✓	✓	0.88
BT-474	GLB1	CMTM7	6	4	✓	✓	✓	0.68
BT-474	CPNE1	PI3	-	-	×	×	×	-
SK-BR-3	TATDN1	GSDMB	118	463	✓	✓	✓	0.29
SK-BR-3	RARA	PKIA	13	10	✓	✓	✓	0.78
SK-BR-3	ANKHD1	PCDH1	-	-	×	×	×	-
SK-BR-3	CCDC85C	SETD3	5	6	✓	✓	✓	0.92
SK-BR-3	WDR67	ZNF704	2	4	✓	×	✓	0.73
SK-BR-3	CYTH1	EIF3H	31	24	✓	✓	✓	0.95
SK-BR-3	DHX35	ITCH	3	4	✓	✓	✓	0.33
KPL-4	BSG	NFIX	20	18	✓	✓	✓	0.90
KPL-4	PPP1R12A	SEPT10	2	6	✓	×	✓	0.65
KPL-4	NOTCH1	NUP214	5	7	✓	✓	✓	0.97
MCF-7	BCAS4	BCAS3	133	212	✓	✓	✓	0.80
MCF-7	ARFGEF2	SULF2	16	40	✓	✓	✓	0.91
MCF-7	RPS6KB1	TMEM49	-	-	×	×	×	-

[1] Crossing reads are the EricScript estimation of the number of reads that supports the discordant alignment.
[2] Spanning reads are the EricScript estimation of number of reads that covers the junction.

this is increasingly true if gene fusions involving middle of exons occur. The large amount of synthetic gene fusions we generated were also used to train an AdaBoost classifier that allows us to assign a reliable probability score to each predicted gene fusion event. The synthetic data generator has been included in the EricScript package: the synthetic data will represent a good resource for new developers when testing their methods. 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 demostrated that EricScript is very reliable in assembling the correct sequence of fusion juctions, allowing for the detection of chimeric events with a resolution of 1bp. The main limitation of our method is the use of a transcriptome instead of a genome reference for mapping reads. Although this option allows us to bring down computational times, it does not enable to discover gene fusions involving unannotated transcribed regions. Recent reports (Cabili *et al.*, 2011) suggest that there are an abundance of unannotated tissue-specific genes: in this case methods such as DeFuse (McPherson *et al.*, 2011) will be more appropriated to screen fusions involving these genes.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

Conflict of interest: none declared.

REFERENCES

Berger, M., Levin, J., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L., Robinson, J., Verhaak, R., Sougnez, C., Onofrio, R., and *et al.* (2010). Integrative analysis of the melanoma transcriptome. *Genome Res*, **20**, 413–27.

Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. (2011). Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes Dev*, **25**, 1915–27.

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I., Nyberg, S., Wolf, M., Borresen-Dale, A., and Kallioniemi, O. (2011). Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biol*, **12**, R6.

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., and Hoeck, W. (2011). Fusionmap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, **27**, 1922–1928.

Iyer, M., Chinnaiyan, A., and Maher, C. (2011). Chimerascan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903–2904.

Jiang, H. and Wong, W. (2009). Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, **25**, 1026–32.

Kent, W. (2002). Blat - the blast-like alignment tool. *Genome Res*, **4**, 656–64.

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

Kinsella, M., Harismendy, O., Nakano, M., Frazer, K., and Bafna, V. (2011). Sensitive gene fusion detection using ambiguously mapping rna-seq read pairs. *Bioinformatics*, **27**, 1068–75.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–9.

Maher, C., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., and Chinnaiyan, A. (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., Pacheco, M., Marra, M., and *et al.* (2011). defuse: An algorithm for gene fusion discovery in tumor rna-seq data. *PLoS Comput Biol*, **7**, e1001138.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*, **7**, 233–45.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**, 621–8.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, **320**, 1344–9.

Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P., Svensson, M., Kitabayashi, N., Moss, B., MacDonald, T., Cao, X., and *et al.* (2011). Discovery of non-ets gene fusions in human prostate cancer using next-generation rna sequencing. *Genome Res*, **21**, 56–67.

Picardi, E., Horner, D., Chiara, M., Schiavon, R., Valle, G., and Pesole, G. (2010). Large-scale detection and analysis of rna editing in grape mtdna by rna deep-sequencing. *Nucleic Acids Res*, **38**, 4755–67.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S., Mungall, K., Lee, S., Okada, H., Qian, J., and *et al.* (2010). De novo assembly and analysis of rna-seq data. *Nat Methods*, **7**, 909–12.

Ruffalo, M., LaFramboise, T., and Koyuturk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**, 2790–6.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D., Rozowsky, J., Tewari, A., Kitabayashi, N., Moss, B., Chee, M., Demichelis, F., Rubin, M., and Gerstein, M. (2010). Fusionseq: A modular framework for finding gene fusions by analyzing paired-end rna-sequencing data. *Genome Biol*, **11**, R104.

Seal, R., Gordon, S., Lush, M., Wright, M., and Bruford, E. (2011). genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*, **39**, D514–9.

Soda, M., Choi, Y., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., and *et al.* (2007). Identification of the transforming eml4-alk fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561–6.

Steidl, C., Shah, S., Woolcock, B., Rui, L., Kawahara, M., Farinha, P., Johnson, N., Zhao, Y., Telenius, A., Neriah, S., and *et al.* (2011). Mhc class ii transactivator ciita is a recurrent gene fusion partner in lymphoid cancers. *Nature*, **471**, 377–81.

Tomlins, S., Rhodes, D., Perner, S., Dhanasekaran, S., Mehra, R., Sun, X., Varambally, S., Cao, X., Tchinda, J., and *et al.* (2005). Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, **310**, 644–8.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511–5.

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**, e164.

Zhang, L., Cheranova, D., Gibson, M., Ding, S., Heruth, D., Fang, D., and Ye, S. (2012). Rna-seq reveals novel transcriptome of genes and their isoforms in human pulmonary microvascular endothelial cells treated with thrombin. *PLoS One*, **2**, e31229.