A novel framework for chimeric transcript detection based on accurate gene fusion model

A novel framework for chimeric transcript detection based on accurate gene fusion model

Francesco Abate, Andrea Acquaviva, Elisa Ficarra, Giulia Paciello, Enrico Macii
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy
Email: francesco.abate, andrea.acquaviva, elisa.ficarra, giulia.paciello, enrico.macii@polito.it

Alberto Ferrarini, Massimo Delledonne
Department of Biotechnology
Università di Verona
Verona, Italy
Email: alberto.ferrarini, massimo.delledonne@univr.it

Simona Soverini, Giovanni Martinelli
Institute of Medical Oncology and Hematology
Università di Bologna
Bologna, Italy
Email: simona.soverini, giovanni.martinelli@unibo.it

***Abstract*—Next generation sequencing plays a key role in the detection of structural variations. Chimeric transcripts are relevant examples of such variations, as they are involved in several diseases. In this work, we propose an effective methodology for the detection of fused transcripts in RNA-Seq paired-end data. The proposed methodology is based on an accurate fusion model implemented by a set of filters reducing the impact of artifacts. Moreover, the methodology accounts for transcripts consistently expressing in the sample under study even if they are not annotated. The effectiveness of the proposed solution has been experimentally validated on of *Chronic Myelogenous Leukemia* (CML) samples, providing both the genes involved in the fusion and the exact chimeric sequence.**

***Keywords*-Next Generation Sequencing, RNA-Seq data, chimeric transcript detection, gene fusions, alternative splicing, deep sequencing analysis, paired-end read**

I. Introduction

The recent advances in cancer research outlined the key role of chimeric transcripts in the characterization of tumor diseases. The success of these results is coupled with the pace of biotechnological field, mainly with the Next Generation Sequencing technology. Specifically, the analysis of RNA-Seq data have been playing a key role in the detection of new fused genes in several disease. In [2], the VAPB-IKZF3 chimera have been found to be involved in the survival in breast cancer cells analyzing RNA-Seq short reads data with a specific bioinformatic pipeline. Moreover, even if the analysis of single long reads has been performed to reveal novel fusion junctions [6], the application of short paired-end reads has been recently demonstrated to provide higher dynamic range and sensitivity in supporting fusion transcripts [8].

In fact, putative chimeric candidates can be discovered analyzing the way the paired-end reads map on the gene fusion boundaries. Figure 1 shows a schematic representation of how the paired-end reads map on two fused genes. Paired-end mate map across the fusion junction in a twofold arrangment: i) Each mate of the read encompasses the junction and maps on a different gene of the fused gene couple. The read is considered as a read *encompassing* the fusion boundary; ii) Alternatively, a single mate of the read overlaps the fusion junction while the corresponding paired-end mate matches one of the two genes involved in the fusion. In this case, the read is considered as *spanning* the fusion junction.

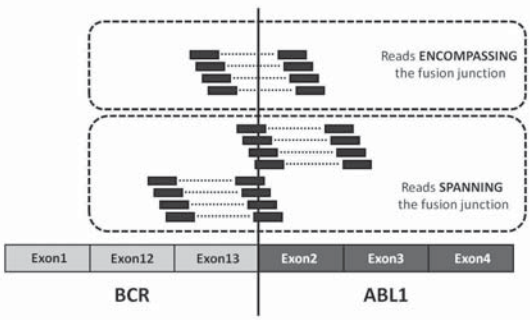


Figure 1. Paired-end reads alignment on a gene fusion junction.

Both encompassing and spanning reads can be used to detect the exact boundary sequence between two fused genes: Encompassing reads allow the production of a first list of gene fusion candidates, while spanning reads are used to detect the exact junction sequence.

However, due to the shortness of the sequenced mates and repetitive regions of the DNA the simple analysis of paired-end reads restricted to the detection of encompassing and spanning set produces a huge list of putative fused gene

candidates. This scenario recalls to the need of an accurate gene fusion model for the detection of those putative chimeric transcripts that most likely fit the model features. In this work, we propose an advanced analysis pipeline for the detection of fusion transcripts through short paired-end reads. In particular, we built an accurate gene fusion model based on recent experimental evidences [2] and we select the most fitting candidates by applying a set of modular filters. Moreover, in order to reduce ambiguous alignments of the reads to isoforms, we perform the analysis on top of a splicing-driven alignment and abundance estimation analysis. This approach allows to account for those transcripts that are consistently expressed in the sample under study, even if they are not annotated. Furthermore, the splicing-driven alignment allows encompassing reads to be mapped more accurately even in presence of proximal splice junctions.

To achieve these targets, the proposed framework leverages upon algorithms such as Cufflinks [10] and TopHat [5], aimed at overcoming RNA-Seq challenges concerning multiple read alignments, novel transcripts discovery and accounting for alternative splicing events.

On this concern, our methodology presents distinguishing features with respect to fusion detection algorithms proposed in the last year [9] [12] [13] [11], in that it integrates these new instruments in a fusion detection framework.

In this paper we report the fusion genes discovered by the proposed framework on experimentally validated biological samples of *Chronic Myelogenous Leukemia* (CML) [14]. Results highlight that the developed methodology, while recognizing the validated fusions, it reduced the final set of predictions and includes fusions involving non-annotated genes.

II. Methods

The flow is mainly composed of two building blocks: *Chimeric Candidates Detection* and *Exact Junction Breakpoint Analysis* (Figure 2). *Chimeric Candidates Detection* aims at providing the list of possible chimeric candidates by detecting and analyzing those reads encompassing putative fusion junctions. *Exact Junction Breakpoint Analysis* relies on the detection of the exact junction breakpoint between two gene candidates through the collection of reads spanning the putative junction breakpoint.

A. Chimeric Candidates Detection

Figure 2 depicts the schematic flow of the *Chimeric Candidates Detection*. This phase is composed of three steps: i) Initial sample alignment to the genome reference; ii) Mapping of read mates to transcripts determined by abundance analysis; iii) Detection of the encompassing reads from the overall set.

The alignment of short RNA-Seq paired-end reads to the reference genome (*Initial Sample Alignment*) is the starting point for determining the list of chimeric candidates. We exploit the capability of TopHat alignment tool [5] to align read fragments on a reference genome considering splicing events and using defalut parameters. At the end of the initial alignment of paired-end reads, both mapped (that include possible encompassing) and unmapped (that include possible spanning) reads are extracted.

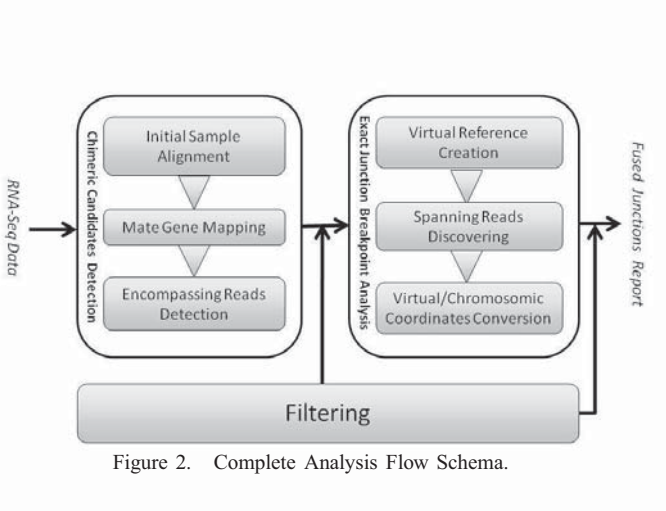


Figure 2. Complete Analysis Flow Schema.

In order to find out candidate genes involved in a fusion event, we need to assign the read location to an annotation file. In the *Mate-Gene Mapping* (see Figure 2) we map each aligned mate on the transcripts detected by transcript abundance analysis by means of Cufflinks [10], thus overcoming the limit of considering only known and annotated transcripts. In fact, analyzing RNA-Seq samples it is possible to reveal new alternative splicing events, novel genes and transcripts that might be neglected in an official annotation file.

The collected set of mapped read mates is analyzed in order to retrieve the subset of reads having the two mates mapping on different genes. At the end of the *Chimeric Candidates Detection* phase the list of possible gene candidates and the set of initially unmapped reads is provided.

On this set of candidates, a cascade of filters is applied to reduce the impact of errors due to the alignment phase as well as artifacts in the preparation of the biological sample [2]. Moreover, ambiguous alignments due to paralogue or homologous regions are taken into account. Related filters will be discussed in Section III.

B. Junction Breakpoint Analysis

Starting from the list of fused candidates previously detected, the scope of the *Exact Junction Breakpoint Analysis* phase, outlined in Figure 2 is to determine the exact junction breakpoint and validate the gene fusion by the alignment of unmapped reads to the putative junction.

From a computational point of view, intron regions cause many mismatches, making alignment programs to fail across

the junction. Splicing discovery programs [4] [5] [7] are aimed at efficiently detecting the exact intron-exon boundary, but due to the considerable computational complexity, they limit their research within a maximum intron size.

To exploit the junction discovery capabilities of splicing detection tools without compromising computational efficiency, *Exact Junction Breakpoint Analysis* adopts a *virtual reference*: 1) For each couple of gene candidates a virtual reference consisting in the concatenation of the two genes is created; 2) A splicing discovery algorithm (i.e. TopHat) is launched on the virtual reference providing as input the initially unmapped reads resulting from the *Chimeric Candidates Detection* phase.

As shown in Figure 2, in order to create a *virtual fusion junction* a *Create Virtual Reference* module automatically retrieves the sequences corresponding to the gene fusion candidates using the coordinates provided by Cufflinks, thus involving both annotated and non annotated transcripts.

The sequences are concatenated and the resulting output represents the *virtual* genome reference of the *virtual* fusion junction. TopHat receives as input the set of unmapped reads and the *virtual* genome reference, resulting from the concatenation of the two gene fusion candidates. TopHat reports all the mapping reads including the spanning end mates. After TopHat alignment, a rearrangement from virtual to chromosomal coordinates is needed, therefore the set of end mates spanning the gene fusion junction is collected and the read coordinates are translated from *virtual* to *genomic coordinates*. End mates spanning the fusion boundaries can be represented as a split read and each chunk maps on a different gene. Thus, the exact points where the first mate chunk ends and the second mate chunk starts represent the exact junction boundary coordinates.

In conclusion, at the end of the *Exact Junction Breakpoint Analysis* for each couple of gene fusion candidates the set of putative junctions, as well as the supporting spanning reads, are reported. However, the detection of spanning reads can be affected by propagation errors due to both alignment limitations and artifacts in the experimental preparation of the sample. For this reason, the resulting junctions are analyzed and filtered depending on how the spanning read maps on each junction. Next section describes the filtering policy applied to improve the accuracy of proposed methodology to discover gene fusions.

III. FILTERS

In this section we deep into details of the filters applied in order to select the chimeric candidates that mostly fit the fusion junction model as described in recent experimental evidences [2]. For clarity sake, we divide the filters into two categories. The former relies on the filters applied to the initial list of candidates resulting from the analysis of the encompassing reads. The latter concerns the analysis of the junction breakpoints detected by the spanning reads discarding those candidates outlining anomalous junction breakpoints.

A. Filters based on encompassing reads analysis

The following filters are applied to the set of putative fused gene candidates resulting from the *Chimeric Candidates Detection* phase. Consequently, they are applied to the set of reads encompassing the two genes implying the candidates.

Ambiguous Encompassing Reads Removal. We recall that gene fusion candidates are detected when two mates of the same read map on different genes. However, it might occur that the same mate maps on multiple transcripts leading to ambiguous encompassing reads detection. The reason is that the lower is the mate length, the higher is the probability of having multiple matches of the same mate. Dealing with short paired-end reads, the probability of having multiple matches in the reference genome is significant, thus mates of the same read might map on multiple genes. In order to remove ambiguous encompassing reads detection we adopt the following strategy: If both the mates of the same read detect multiple couples of gene fusion candidates but both the mates also match on the same gene, the read is discarded. By looking at Figure 3 two mate ends of the same ReadA individuate two set of genes determining multiple couples of gene fusion candidates (GeneA-GeneB and GeneB-GeneC). In the case shown in figure, the two gene sets share a gene. So, they are not considered because it is highly probable that the two mates actually belong to the same genes. Therefore, if the two gene sets have no gene in common, they do not map on the same gene and all the possible combinations of genes belonging to the two sets are considered as possible candidates.

Abnormal Inner Size Filter. We focus now the discussion on the second type of filter which implements a strategy similar to [9], where a distribution of the inner distance between encompassing reads is estimated. Those candidates presenting the inner distance that is outliers respect to the fragment inner distance mean are removed. The inner distance is computed through consensus regions as depicted in Figure 4. In this work we also propose a new and extended implementation that takes into consideration a recent observation in [2] about the asymmetry in the alignment. This is recognized to be a feature of artifacted chimeric transcripts.

This filter looks at consensus regions made by encompassing reads on the candidate genes. The length of these regions is computed (excluding possible gaps in between) as shown in Figure 4. If one of the two regions, for instance the one related to candidate A in Figure 5, is much larger than the corresponding consensus region of the

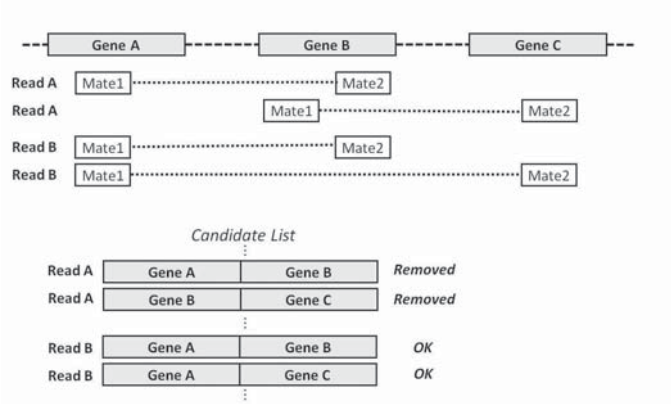


Figure 3. Candidates selection in case of multiple mismatches. The Mate1 of the ReadA maps on the GeneA and GeneB, whereas the Mate2 of the ReadA maps on GeneB and GeneC. The ReadA individuates two set of genes with the GeneB in common, thus the read is discarded. Conversely, Mate1 of ReadB maps on GeneA whereas Mate2 of ReadB maps on GeneB and GeneC. Therefore, all genes in the gene set are distinct and ReadB individuates two gene fusion candidates.

candidate B, the couple A-B is discarded.

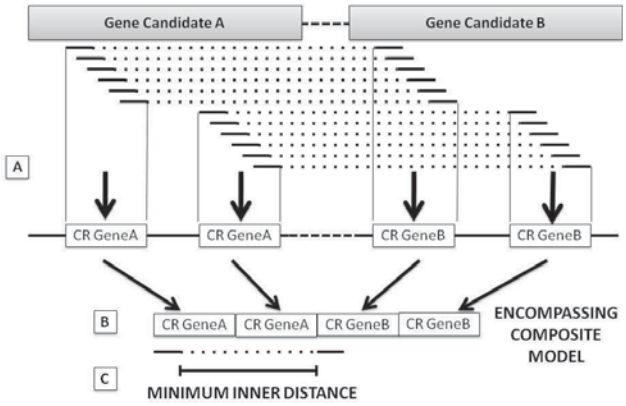


Figure 4. Consensus Regions and Inner Distance Computation.

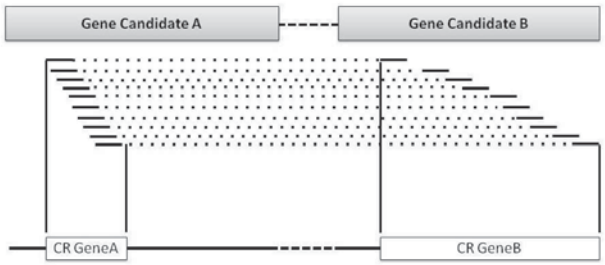


Figure 5. Asymmetric Abnormal Fusions.

Homologous Sequence Filter. After filtering the reads involved in artifacts and alignment errors, another set of filters on chimeric candidates is performed. Candidates supported by a percentage of ambiguous reads with respect to the total number of reads are discarded. Ambiguous reads are caused by short or long homologous sequences in the reference genome. Fusion detection analysis is affected because the mate pairs that, without homologous sequences, would match on the same gene, match discordantly on two distinct but similar genes, thus creating fake encompassing reads. Homologous regions may be due both to the presence of paralogue genes that share long sequence regions and to the presence of short similar sequences.

The *Homologous Sequence Filter* implements two different policies for both cases. Concerning the long homologous sequences due to paralogue genes a filter that query TreeFam [3] database has been implemented. For short homologous sequences, we apply a strategy similar to what proposed in [9], where read mates encompassing the fusion candidates are extracted and reversely mapped on the same genes.

Figure 6 shows the main idea behind the adopted algorithm. If the mapping is confirmed it means that the read encompasses the candidates due to a homologous subsequence. The candidate is then discarded in the case the ratio between the number of ambiguous reads and total number of encompassing reads is greater than a user defined threshold.

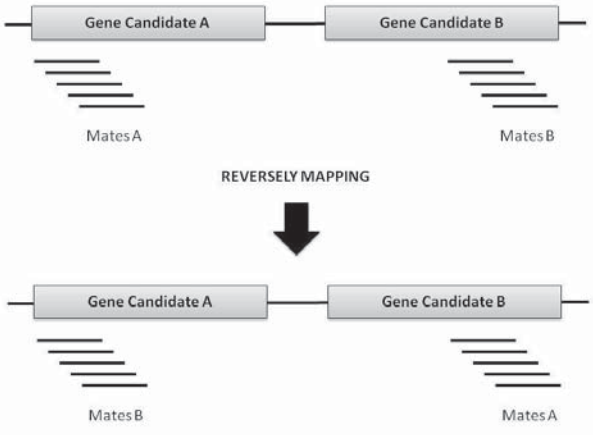


Figure 6. Short Homologous Sequence Filter.

Additional Filters. The remaining filters look at gene candidate distance and number of supporting encompassing reads. In particular, fusions occurring between genes closer than a user defined threshold are filtered out by the *Neighbor Candidate Filter*, as they are considered instances of transcriptional readthroughs [2]. Finally, since both alignment bias and biological sample preparation

artifacts produces false fusion candidates that are typically supported by a small number of encompassing reads, chimeric candidates having the number of encompassing reads below a user defined threshold are filtered out by the *Supported Candidates Thresholding Filter*. The threshold value depends on the coverage of the overall sequencing experiment and adopted protocol.

B. Filters based on spanning reads analysis

The *Exact Junction Breakpoint Analysis* provides a list of putative junctions boundaries between two fused genes. A selection is performed at this stage by looking at the distribution of the reads spanning the junction, to reveal possible artifacts. Therefore, we apply some filters in order to remove all the artifact junctions from the resulting set and to make junctions list more accurate.

Floating Fragment Removal Filter. It might occur that the same read mate maps on the putative junction in multiple ways. In fact, some subsequences of the gene sequence might be homologous and consequently some small fragments of the read mate matche the candidate gene in multiple places of the sequence. Thus, these fragments float on multiple places of the gene sequence and the accuracy of their mapping may be compromised. Furthermore, when this scenario occurs, TopHat reports a distinct read mate instance for each multiple match. However, this does not lead to a realistic count of the number of read mates supporting the junction.

To address this issue, we propose *Floating Fragment Removal Filter*, that removes all the small floating fragments of the read mate sequence mapping on multiple places of the reference gene. Figure 7 depicts an example where the second mate is characterized by fragments mapping on different locations. Specifically, this filter detects and preserves all those read mate subsequences mapping the reference in the same region. In this way, only those read portions that are highly probable to be correctly mapped on the reference sequence are considered to support the putative junction. Moreover, as only the commonly mapped subsequences are preserved, it is pointless to report multiple instances of the same read mate, therefore the mate is considered as unique.

PCR Artifacts Removal Filter. A second filter, named *PCR Artifacts Removal Filter*, is based on the observation that PCR amplification might cause false putative junctions [2]. Reads mapping exactly to the same position are likely due to an artifact originated by amplification and sequencing of the same initial fragment repeated more times. As a result, multiple identical reads are considered as a single one and the fusion supported by those reads will

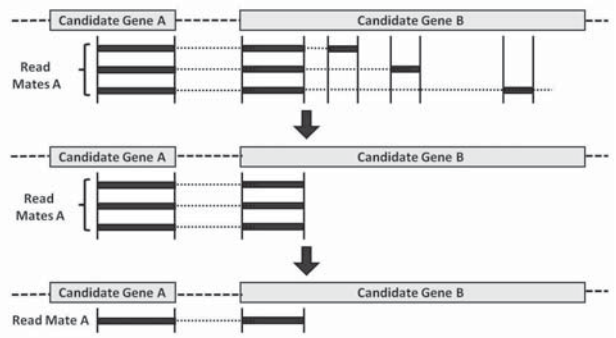
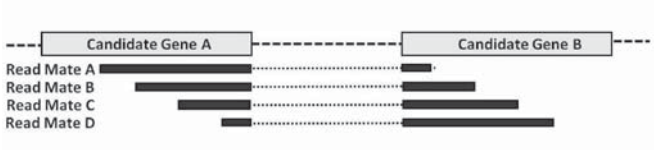


Figure 7. Floating Fragment Removal Filter.

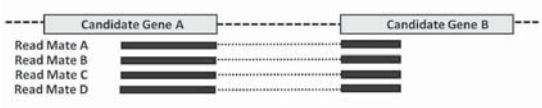
be discarded unless other supporting spanning reads will be found.

Real fusions are characterized by a ladder-like pattern of the spanning reads supporting the junction (Figure III-B). Conversely, false positive junctions due to PCR amplification artifacts lack this pattern and all the short read mates spanning the junction either map on the same position or are one or two bases shifted (Figure III-B). The *PCR Artifacts Removal Filter* removes all those putative junctions lacking the ladder-like pattern.

A final filtering is performed on: i) Candidate fusions supported by a number of spanning reads lower than a user defined threshold (*Supported Junctions Thresholding Filter*); ii) Coherency with encompassing reads. In particular, the latter is based on the observation that, in presence of a genuine gene fusion, the genomic coordinates of the set of encompassing reads must be adjacent or in some cases overlapped to the location of the spanning reads. Therefore, consensus of the final set of both encompassing and spanning reads is created. If the coordinates of the corresponding locations are either adjacent of overlapped the fusion candidate is confirmed otherwise it is discarded.



(a) Read mates correctly spanning the junction



(b) Read mates spanning the junction due to PCR Artifact

Figure 8. PCR Artifact Removal Filter.

Table I
FUSIONS PREDICTED ON *Chronic Myelogenous Leukemia* SAMPLES.

Lib.	[#] Reads (Mlns)	Read Len.	Frag. Mean	Frag. Stdev	Total Fus.	Inter Chr.	Intra Chr.
s_4	20	75	212	16	2	2	0
s_7	32	75	225	19	4	3	1
s_8	29	75	229	22	10	9	1

IV. RESULTS

We evaluated the effectiveness of the proposed pipeline of analysis in detecting chimeric transcripts analyzing three samples of CML progression from a Philadelphia chromosome-positive (Ph+) *Chronic Myeloid Leukemia* (CML) patient [14] For CML samples we report a complete characterization including the effects of filters. For CML samples we report the number of detected fusions and the details of the applied filters on the initial set of chimeric candidates. The RT-PCR analysis on all the CML samples presents the well-known chromosomal translocation between BCR and ABL1 genes and in all the samples the exact sequence of the chimeric fusion has been detected. All the CML samples are 75 length paired ends produced by an Illumina Genome Analyzer II. For all the reported analysis the *GRCh37/hg19 Feb. 2009* assembly of the human genome and the GRCh37 file for annotations from Ensembl have been adopted. Table I reports the statistics concerning all the CML samples and details about the number of total fusions detected. In the last column, the number of intra chromosomic fusions are shown. Filter parameters used for these runs have been set as follows. Minimum supporting reads: 8; Inner distance threshold: 400bp.

Under the hypothesis that the scientist is not interested in adjacent fused genes [2], that can be detected by classical splicing detection tools, we set the *Neighbor Candidates Filters* with 500000 bp thresholds and this caused most of the revealed fusions to be inter-chromosome.

A. Material and Samples Preparation

Three samples from a Philadelphia chromosome-positive (Ph+) CML patient were previously sampled (sample s_4, s_7 and s_8) [14]. The patient was diagnosed with Ph+ p210BCR-ABL-positive CML by chromosome banding analysis. The samples were tested for rearrangements between BCR and ABL genes by reverse transcription-polymerase chain reaction (RT-PCR) [16].

RNA-Seq libraries (one per sample) were prepared using the mRNA-Seq 8 sample preparation kits following manufacturer instructions. We modified the gel extraction step by dissolving excised gel slices at room temperature to avoid underrepresentation of AT-rich sequences [15]. Library quality control and quantification was performed with a Bioanalyzer Chip DNA 1000 series II (Agilent). Libraries were sequenced on an Illumina genome analyzer II following manufacturer instructions and 75 bp paired-end reads were obtained.

Table II
EFFECTS OF APPLIED FILTERS. THE REPORTED PERCENTAGES ARE COMPUTED AS THE RATIO BETWEEN THE NUMBER OF FILTERED CANDIDATES AND THE NUMBER OF CANDIDATES FILTERED BY THE PREVIOUS FILTER.

Lib.	Initial Cand.	Supp. Cand. Thrs.	Naming Incoher. (*)	Neigh Cand.	Abnormal Inner Dist.	Finally Not Filtered	[%] Not Filt.
s_4	24337	87%	8%	6%	34%	1754	7%
s_7	86552	94%	13%	9%	36%	2482	3%
s_8	122931	95%	19%	8%	41%	2791	2%

*Naming incoherencies are detected when the same gene name share different gene identifier in Cufflinks annotation. Thus two apparently different gene are actually a single gene.

Table III
HOMOLOGOUS SEQUENCE FILTER ON CML SAMPLES.

Library	Total Candidates	Selected [%]	Removed [%]	Ratio Threshold	Mismatch Allowed
s_4	1754	4	96	0,0	25
s_7	2482	3	97	0,0	25
s_8	2791	8	92	0,0	25

B. Filtering Effects

We detail the effects of the various filters applied both to the candidate fusions and to the spanning reads. We report filtering results of the CML samples. Table II and III shows the effect of filters on candidate fusions (Section II-A) while Table IV refers to spanning reads filtering. Numbers in the tables report the percentage of candidates removed with respect to the previous filter in the cascade.

Effects of filters on candidate fusions. The considerable number of initial candidates detected in the first phase by discordant reads mapping (up to 122931 in *s_8* sample) is consistently reduced through the pipeline of filters shown in Table II. A large fraction of candidates is discarded because was not supported by a sufficient number of encompassing reads (see third column in Table II). Moreover, 34%-41% of putative fusions with a sufficient number of encompassing reads has been removed because of abnormal inner size and asymmetry in consensus regions.

Table III details the effect of *Homologous Sequence Artifacts Filter*. Because of its large computational cost due to the reverse remapping of the encompassing reads, this filter has been applied as a final step on a reduced set of candidates. This filter was very selective, leaving 3%-8% of putative candidates for the following spanning analysis phase.

Effects of filters on junction artifacts. Both alignment bias and biological artifacts due to PCR amplification might cause the detection of false putative junctions. In order to mitigate the negative effects of these events on the chimeric transcript analysis, the filters described in Section III-B

Table IV
EFFECTS OF THE FILTERS ON THE PUTATIVE JUNCTIONS.

Sample	Floating Fragm. Rem. Filter		PCR Artifact Rem. Filter		Less Supported Junction Rem. Filter	
	[#]	[%]	[#]	[%]	[#]	[%]
s_4	214	0	178	17	82	62
s_7	227	0	206	9	125	45
s_8	427	0	339	21	263	39

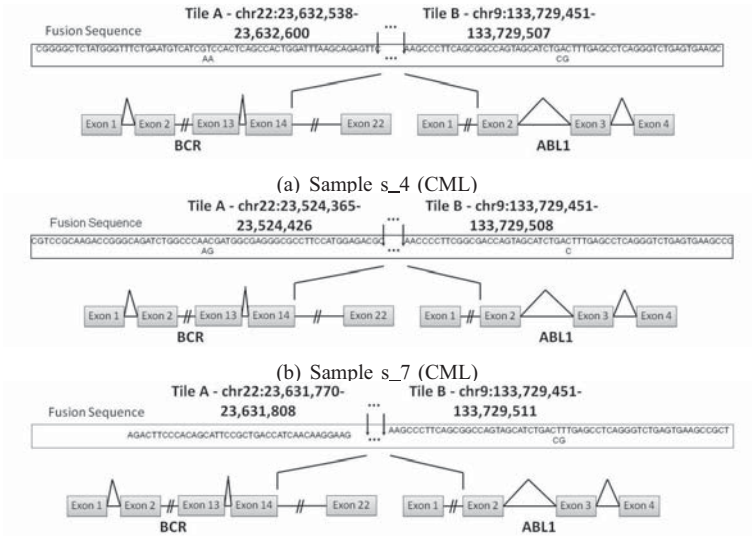
are applied during the *Exact Junction Breakpoint Analysis* phase. Table IV reports the effect of the application of the filters on the initial number of putative junctions detected for each sample. The initial number of junctions (i.e. spanning reads) is in general larger than the candidates resulted from the first encompassing analysis phase, since each candidate has multiple spanning reads associate to it.

The *Floating Fragment Removal Filter* does not reduce the number of the initial putative junctions. However, it plays a fundamental preliminary role for the following *PCR Artifacts Removal Filter*. In fact, the floating fragments cause false ladder-like patters that are actually replicas of the same reads (see Figure 7). The *Floating Fragment Removal Filter* removes the floating fragments and it allows a more accurate detection of PCR artifacts. Therefore, the *PCR Artifacts Removal Filter* removes from the 9% to 20% of false putative junctions and the number of junctions ranges in the best case (sample *s_8*) from 427 to 339. Moreover, the removal of the floating fragments makes in some cases to considerably decrease the number of reads spanning across the putative junction (See Figure 7). Consequently, the *Supported Junctions Thresholding* filter is more effective after applying *Floating Fragment Removal Filter* and the reduction spans from 39% to 62%.

C. Exact Junction Discovery Details

Figure 9 depicts the results of the *Exact Junction Breakpoint Analysis* applied to CML samples. A group of reads is mapped onto the reference genome spanning the fusion boundary between the BCR and ABL1 genes. This chimera has been validated through RT-PCR analysis. The spanning reads present reported in Figure 9 show a ladder-like pattern across the junction boundary according to the junction model we use in this work. The exons involved in the fusion are exon 14 (*s_4*), exon 1(*s_7*) and exon 12 (*s_8*) of BCR and exon 2 of ABL1.

Fusions involving non-annotated genes. Being based on transcript expression analysis instead of annotated genes, fusions involving non-annotated genes have been detected. The analysis of sample s_8 reveals that 3 on 10 fusions involved non-annotated transcripts. This result is relevant for the detection of new aberrant modifications in the gene regulation, which is one of the main targets of next generation sequencing analysis.



(a) Sample s_4 (CML)

(b) Sample s_7 (CML)

(c) Sample s_8 (CML)

Figure 9. Junction Boundary Detection.

V. CONCLUSION

In this paper we presented a new methodology for the detection of chimeric transcripts in RNA-Seq paired-end data. The main contribution of the proposed approach is the capability of selecting chimeric candidates fitting an accurate fusion transcript model. Moreover, the flow is able to detect gene fusions involving both annotated and non annotated transcripts. Reads alignment phase takes into account splicing events directly derived from experimental data, enhacing the overall alignment accuracy.

The proposed strategy has been applied to real *Chronic Myelogenous Leukemia* samples and the detected fusions have been validate through RT-PCR analysis. These results highlight the effectiveness of the proposed approach in the detection of novel fused transcript discovery.

REFERENCES

[1]

[2] Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. *Identification of fusion genes in breast cancer by paired-end RNA-sequencing.* Genome Biology, 2011 January.

[3] Li H, Coghlan A, Ruan J, Coin L J, Hrich J K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R. *TreeFam: a curated database of phylogenetic trees of animal gene families.* Nucleic Acids Research, 2006 January.

[4] Bryant D W Jr, Shen R, Priest H D, Wong WK, Mockler T C. *Supersplat–spliced RNA-seq alignment.* Bioinformatics, 2010 January.

[5] Trapnell C, Pachter L, Salzberg S L. *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009 May .

[6] Maher C A, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan A M. *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009 March.

[7] Ameur A, Wetterbom A, Feuk L, Gyllensten U. *Global and unbiased detection of splice junctions from RNA-seq data.* Genome Biology, 2010 March.

[8] Maher CA, Palanisamy N, Brenner J C, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette T R, Grasso C, Yu J, Lonigro R J, Schroth G, Kumar-Sinha C, Chinnaiyan. *Chimeric transcript discovery by paired-end transcriptome sequencing.* PNAS, 2009 July.

[9] Sboner A, Habegger L, Pflueger D, Terry S, Chen D Z, Rozowsky J S, Tewari A K, Kitabayashi N, Moss B J, Chee M S, Demichelis F, Rubin M A, Gerstein M B. *FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.* Genome Biology, 2010 October.

[10] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren M J, Salzberg S L, Wold B J, Pachter L. *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature Biotechnology, 2010 May.

[11] Daehwan Kim and Steven L Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts*, Genome Biology, 2011. http://tophat-fusion.sourceforge.net.

[12] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. *deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data.* PLoS Computational Biology, 2011 May.

[13] Li Y, Chien J, Smith D I, Ma J. *FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq.* Bioinformatics, 2011 May.

[14] Soverini S, Score J, Iacobucci I, Poerio A, Lonetti A, Gnani A, Colarossi S, Ferrari A, Castagnetti F, Rosti G, Cervantes F, Hochhaus A, Delledonne M, Ferrarini A, Sazzini M, Luiselli D, Baccarani M Cross N C P, Martinelli G. *IDH2 somatic mutations in chronic myeloid leukemia patients in blast crisis.* Leukemia:official journal of the Leukemia Society of America, 2011 January.

[15] Michael A. Quail, Harold Swerdlow, Daniel J. Turner *Improved Protocols for the Illumina Genome Analyzer Sequencing System.* Current Protocols in Human Genetics, 2009 July.

[16] Dongen et al., *Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets.* Leukemia, 1999 January.