The analysis of RNA-Seq data revealed to be particularly effective supporting fusion transcripts (Maher demonstrated to provide higher dynamic range and sensitivity in However, the application of short paired-end reads has been recently long reads has been performed to reveal novel fusion junctions. (Edgren been found to play a key role in the survival in breast cancer cells involved in several diseases. By applying RNA-Seq short reads, novel fusion genes and in particular the VAPB-IKZF3 chimera have been found to play a key role in the survival in breast cancer cells (Edgren et al 2011). In Maher et al (2009), the analysis of single long reads has been performed to reveal novel fusion junctions. However, the application of short paired-end reads has been recently demonstrated to provide higher dynamic range and sensitivity in supporting fusion transcripts (Maher et al 2009). In paired-end
gene fusion model. The model, based on validated experimental evidence, is implemented by Bellerophontes through a set of modular filters.

- It integrates a splicing-driven alignment and abundance estimation analysis, leading to a more accurate set of reads supporting the junction discovery because it reduces ambiguous assignments of reads to isoforms and allows the detection of novel fused transcripts. Furthermore, this approach allows to account for those transcripts that are consistently expressed in the sample under study, even if they are not annotated.

- The full pipeline has been developed on top of a splicing-driven alignment. As a result, encompassing reads are mapped more accurately even in presence of proximal splice junctions. This enhances the accuracy of reads supporting fusion candidate detection and junction sequence discovery.

To achieve these targets, Bellerophontes leverages upon algorithms such as Cufflinks \cite{Trapnell2010}, TopHat \cite{Trapnell2009}, and TopHat-Cufflinks \cite{Trapnell2010}, aimed at overcoming RNA-Seq challenges concerning multiple read alignment, novel transcript discovery and accounting for alternative splicing. In addition, it exploits recent researches about the pattern of reads mapping across fusion breakpoints \cite{Edgren2011}, enabling a more accurate model of the fusion junction. On this concern, Bellerophontes presents distinguishing features with respect to fusion detection tools proposed in the last year \cite{Li2010, TopHat2010, Sboner2010}, http://tophat-fusion.sourceforge.net, in that it integrates these new instruments in a fusion detection software framework.

In this article, we report the fusion genes discovered by the proposed framework on experimentally validated biological samples of chronic myelogenous leukemia (CML) \cite{Soverini2010}, and on public NCBI datasets of validated fusions. We also performed a comparative analysis with an alternative state-of-art approach \cite{McPherson2011} on the same datasets. The results highlight that Bellerophontes, while recognizing the validated fusions, reduced the final set of predictions and includes fusions involving non-annotated genes.

2 METHODS

2.1 Procedure and algorithms

This section first focuses on the computational infrastructure of the proposed tool and, then reports details about the application of the tool to real data analysis. The flow is mainly composed of two building blocks: ‘chimeric candidates detection’ and ‘exact junction breakpoint analysis’ (Fig. 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.

2.1.1 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 and (iii) detection of the encompassing reads from the overall set.

RNA-Seq data analysis framework for chimeric transcripts

**Fig. 2.** Complete analysis flow schema of the proposed tool. (a) ‘chimeric candidates detection’ and (b) ‘exact junction breakpoint analysis’

**Initial sample alignment.** The starting point for determining the list of chimeric candidates is the alignment of short RNA-Seq paired-end reads to the reference genome (initial sample alignment). Most of fusion detection tools adopt Bowtie \cite{Langmead2009} for aligning paired-end reads to the reference genome. Conversely, we exploit the capability of TopHat alignment tool \cite{Trapnell2009} to align read fragments on a reference genome considering splicing events. TopHat reports variable length alignments due to the presence of splicing junction breakpoints. Consequently, the framework has to take into account aligned spliced alignments to the reference genome instead of fixed length reads. 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.

**Mate-gene mapping.** 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 Fig. 2), we map each aligned mate on the transcripts detected by transcript abundance analysis by means of Cufflinks \cite{Trapnell2010}, 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. This is relevant in the context of chimeric transcripts, which are unpredictable events. In this context, a fusion possibly involves non-annotated genes or genes showing intron retentions. Considering only annotated gene isoforms, those reads encompassing the non-annotated region would be discarded and thus the fusion would not be detected.

**Discordant mates detection and candidate filtering.** 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 are 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 \cite{Edgren2011}. Moreover, ambiguous alignments due to paralogous or homologous regions are taken into account. Related filters will be discussed at the end of this section. As regards filters on artifacts, they are aimed at discarding: (i) reads that detect multiple couples of gene fusion candidates that, besides their discordant matching, they also have both mates on the same gene and (ii) reads showing an abnormal inner size (computed as described in Fig. 3) between the sequenced ends, or asymmetry in the alignment of the mates encompassing a fused gene (an example is shown in Fig. 3).

We focus now the discussion on the second type of filter, namely, the ‘abnormal inner size filter’, which implements a strategy similar to Sboner et al. \cite{Sboner2010}, where the distribution of the inner distances in the sample is computed and outliers are removed based on a threshold. In this work, this has been set to 400 bp. In Bellerophontes, inner distance is computed through consensus regions as depicted in Figure 3. This is a minimum inner distance,
larger than the sum of the consensi determined through encompassing reads. As such, the inner distance is in general because they may not be complete, as reads spanning the junction in this phase of the pipeline are not considered. As such, the inner distance is in general larger than the sum of the consensus determined through encompassing reads.

In this work, we also propose a new and extended implementation that takes into consideration a recent observation in [Edgren et al., 2011] about the asymmetry in the alignment. This is recognized to be a feature of artifactual 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 3. If one of the two regions, for instance the one related to Candidate A in Figure 3a, is much larger than the corresponding consensus region of the Candidate B, the couple A-B is discarded.

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 (see Table 1 in Section 3 for details). 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 artifacts filter’ implements two different policies for both cases. Concerning the long homologous sequences due to paralogue genes, a filter that query TreeFam [Ameur et al., 2009] database has been implemented. For short homologous sequences, we apply a strategy similar to what proposed in [Chen et al., 2011], where read mates encompassing the fusion candidates are extracted and reversely mapped on the same genes.

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 [Edgren et al., 2011]. Finally, since both alignment bias and biological sample preparation artifacts produce 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.

2.1.2 Junction breakpoint analysis

Starting from the list of fused candidates previously detected, the scope of the exact junction breakpoint analysis phase, outlined in Figure 4, 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, the problem of finding gene fusion junctions can be considered an extension of the detection of splicing events involving exons couples belonging to different genes and chromosomes. Splicing events introduce a considerable level of complexity in the analysis of RNA-Seq fragments. Intron regions cause many mismatches, making alignment programs to fail across the junction. Splicing discovery programs [Chen et al., 2011; Bhat et al., 2011; Trapnell et al., 2012] are aimed at efficiently detecting the exact intron–exon boundary. 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: (i) for each couple of gene candidates, a virtual reference consisting in the concatenation of the two genes is created and (ii) 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 4c, 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. The corresponding sequence is retrieved from the reference in ‘UCSC Genome Browser database’ [Fujita et al., 2010]. Being based on Cufflinks coordinates, the sequences retrieved on the reference may correspond to annotated or non-annotated genes.

The sequences are then concatenated and the resulting file 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. The result of TopHat is a file containing all the mapping reads including the spanning end mates. After TopHat alignment, a rearrangement from virtual to chromosomal coordinates is needed.

Once the set of end mates spanning the gene fusion junction is collected and the read coordinates are translated from virtual to genomic ‘coordinates’, it is possible to exactly determine the boundary junction among the two gene candidates. As shown in Figure 4d, end mate reads spanning the fusion junction can be represented as a split read and each chunk maps on a different gene section. It is worth noting that all the spanning mate chunks are spaced by a gap in the genome reference. 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...
Fig. 5. Schematic representation explaining how spanning mates reveal exact genome coordinates of gene fusion junction. Spanning mates are indicated as black solid lines separated by a dashed line in correspondence of the junction breakpoint.

Fig. 6. Floating fragment removal filter. Some subsequence of the same Read Mate A maps differently on the two Candidate Gene A and B. Specifically a small fragment floats along the Candidate Gene B. The floating fragment removal filter removes the floating fragments considering as a valid mapping only the subsequences commonly mapped to the reference. Finally, only a single instance of read mapping is reported.

Fig. 7. PCR artifact removal filter. (a) The schema of a group of read mates spanning a genuine fusion junction. The short reads present a ladder-like pattern and they all map on distinct points of the candidate genes involved in the fusion. (b) An example of a group of read mates mapping a false breakpoint junction. Short reads are more overlapped as all the reads share a similar start and end mapping locations.

2.1.3 Spanning read analysis and junction filtering. 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 list and to make junctions list more accurate.

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 match the candidate gene in multiple places of the sequence. Thus, these fragments float on multiple places of the genome 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 6 on the upper part, shows 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 (see the middle part of Fig. 6). 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 a unique (see bottom part of Fig. 6).

A second filter, named PCR artifacts removal filter, is based on the observation that PCR amplification might cause false-putative junctions from the resulting set and to make junctions list more accurate.

2.2 Real data analysis

2.2.1 Materials and sample preparation. We present result about the detection of chimeric transcripts on three datasets at different stages of CML progression from a Philadelphia chromosome-positive (Ph+) CML patient (sample s_A, s_B, s_C, and s_D). The patient was diagnosed with Ph+ p210(BCR-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). The samples were processed for complementary DNA (cDNA) synthesis and PCR amplification using primers specific for the translocation breakpoint junction. Short reads present a ladder-like pattern and they all map on distinct points of the candidate genes involved in the fusion. Short reads are more overlapped as all the reads share a similar start and end mapping locations.

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

The second dataset we used for evaluation is a public sets of RNA-Seq data including information about RT-PCR validated fusions. The dataset was obtained from NCBI database (http://www.ncbi.nlm.nih.gov/sra) under submission numbers SRA009053 and SRA040160.
Table 1. Fusions predicted with Bellerophontes on chronic myelogenous leukemia samples

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

Table 2. Fusions predicted on publicly available RNA-Seq data

<table>
<thead>
<tr>
<th>Library</th>
<th>Reads [#] (millions)</th>
<th>Read length</th>
<th>Fragment length</th>
<th>Validated predicted fusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>018259</td>
<td>14</td>
<td>50</td>
<td>500</td>
<td>1</td>
</tr>
<tr>
<td>018260</td>
<td>16</td>
<td>50</td>
<td>500</td>
<td>2</td>
</tr>
<tr>
<td>018261</td>
<td>16</td>
<td>50</td>
<td>500</td>
<td>1</td>
</tr>
<tr>
<td>018265</td>
<td>8</td>
<td>50</td>
<td>500</td>
<td>1</td>
</tr>
<tr>
<td>018266</td>
<td>15</td>
<td>50</td>
<td>500</td>
<td>4</td>
</tr>
<tr>
<td>018267</td>
<td>15</td>
<td>50</td>
<td>500</td>
<td>2</td>
</tr>
<tr>
<td>018269</td>
<td>15</td>
<td>50</td>
<td>350</td>
<td>3</td>
</tr>
<tr>
<td>NCIH660</td>
<td>7</td>
<td>50</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

All the library identifiers, with exception of the last row, refer to the accession number reporting the SRR prefix in the NCBI databank.

2.2.2 Fusions detected on real samples

For both CML samples and NCBI datasets, we report the number of detected fusions. For all the reported analysis Bellerophontes ran with GCCCh37/hg February 19, 2009 assembly of the human genome, while we used GRCh37 file for annotations from Ensembl.

Table 2 reports the statistics concerning all the CML samples and details about the number of total fusions detected. In all the samples, Bellerophontes detects the exact sequence of the chimeric fusion. In the last column, the number of intra chromosomal fusions are shown. Filter parameters used for these runs have been set as follows. Minimum supporting reads: 8; neighbor candidates filters: 500 000 bp; inner distance threshold: 400 bp.

Under the hypothesis that the scientist is not interested in adjacent fused genes (Edgren et al., 2011), that can be detected by classical splicing detection tools, we set the neighbor candidates filters with 500 000 bp thresholds and this caused most of the revealed fusions to be inter-chromosome.

Table 2 highlights the capability of Bellerophontes concerning the detection of validated fusions on a published dataset. These samples have lower coverage (at most 16 million reads) and smaller read length (50bp) compared with CML data (see Table 1 for comparison). All the 14 fusions validated in the seven samples of melanoma cells (Edgren et al., 2011) and the TMPRSS2-ERG fusion (causally linked to prostate cancer) in NCIH660 (Sboner et al., 2010) have been successfully detected. For the analysis of NCBI samples, some filter constraints have been set to be less stringent with respect to CML samples, because of the specific characteristics of NCBI dataset in terms of coverage and presence of readthroughs. In particular, the threshold on the number of encompassing in the supported candidate thresholding filter was set to 2, since a lower coverage may cause a smaller number of supporting encompassing reads (see Table 2). Furthermore, the threshold for the neighbor candidate filter was set to 50 bp to account for the larger number of readthroughs in this dataset.

Fig. 8. Junction boundary detection. Results of the exact junction breakpoint analysis applied to CML samples. The fusion sequence corresponding to BCR and ABL1 genes is reported as well as the group of reads spanning the junction boundary. The exons involved into the fusion are also indicated in correspondence of the fusion boundary points.

2.2.3 Exact junction discovery details

Figure 8 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 reported in Figure 8 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 13 (s_8) of BCR and exon 2 of ABL1.

Fusions involving non-annotated genes

Using transcript expression analysis instead of annotated genes, Bellerophontes is able to detect fusions involving non-annotated genes. The analysis of Sample s_8 reveals that 3 of 10 fusions involved non-annotated transcripts. For instance, Bellerophontes identifies a fusion involving EWSR1 gene (chr22:29664273-29669806) and a not annotated gene (namely, CufflinksNovelGene chr14: 36350337-36350395). This feature 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.

3 RESULTS

In this section, we report details about the effects of filters and the results about a comparative analysis with another recently published
We detail the effects of the various filters applied both to the gene fusions and to the spanning reads. We report filtering results of only the CML samples, but similar considerations can be drawn for NCBI samples. Table 3 shows the effect of filters on candidate fusions (Section 2.1.1) while Figure 9 refers to spanning reads filtering. Numbers in the tables report the percentage of candidates removed with respect to the previous filter.

### 3.1 Filtering effects

#### 3.1.1 Effects of filters on candidate fusions

The considerable number of initial candidates detected in the first phase by discordant read mapping (up to 122,931 in s_8 sample) is consistently reduced through the pipeline of filters shown in Table 3. A large fraction of candidates is discarded because it was not supported by a sufficient number of encompassing reads (see third column in Table 3). Moreover, 32%–37% of putative fusions with a sufficient number of encompassing reads has been removed because of abnormal inner size and asymmetry in consensus regions.

Because of its large computational cost due to the reverse remapping of the encompassing reads, the homologous sequence artifacts 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.

#### 3.1.2 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 2.1.3 are applied during the exact junction breakpoint analysis phase. Figure 9 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 of the tool, since each candidate has multiple spanning reads associated to it. The floating fragment removal filter does not reduce the number of the initial putative junctions. However, it plays a fundamental role for the following PCR artifacts removal filter. In fact, the floating fragments cause false ladder-like patterns that are actually replicas of the same reads (see Figure 9). 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 9). Consequently, the ‘supported junctions thresholding’ filter is more effective after applying floating fragment removal filter and the reduction spans from 39 to 62%.

### 3.2 Comparative analysis

To highlight that the proposed approach provides results comparable with state-of-art tools while it introduces original features improving fusion discovery, we compared Bellerophontes analysis with deFuse approach McPherson et al. (2011), which has been used to discover gene fusions in tumour samples Susici et al. (2010).

Table 4 reports the number of fusions predicted with Bellerophontes and deFuse on CML samples. This table also shows the number of spanning reads (i.e. <8). Those fusions are discarded by Bellerophontes because of the supported candidate thresholding filter, whose threshold was set to 8.

The smaller number of predicted fusions is also due to the usage of Cufflinks to map the discordant read mates. Indeed, only candidates with a minimum level of abundance are considered, reducing the set of possible fused genes to be evaluated in the successive stages of the pipeline. Indeed, we observed that most of the candidates provided by deFuse have been discarded by Bellerophontes because the related genes are filtered out by Cufflinks with default parameters, due to their low abundance. Note that, if we are interested in detecting junctions between poorly expressed candidates, it is possible to run Cufflinks with less restrictive
The proposed framework implements a new analysis pipeline that
To quantify the impact of Cufflinks, we performed experiments with
Table 4. Comparison between Bellerophontes and deFuse CML Data

<table>
<thead>
<tr>
<th>Lib.</th>
<th>deFuse</th>
<th>Bellerophontes</th>
<th>deFuse</th>
<th>Bellerophontes</th>
<th>deFuse</th>
<th>Bellerophontes</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_4</td>
<td>21</td>
<td>2</td>
<td>15</td>
<td>14</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>s_7</td>
<td>46</td>
<td>4</td>
<td>48</td>
<td>36</td>
<td>46</td>
<td>35</td>
</tr>
<tr>
<td>s_8</td>
<td>52</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

parameters. Note also that the set of candidates given by deFuse
is not a superset of our candidates, this means that we find some
candidates, that deFuse does not find, having a relevant expression.
To quantify the impact of Cufflinks, we performed experiments with
relaxed parameters. We observed an increasing number of fusions
in common with deFuse, up to 85%. We do not state that these
additional fusions are false positives. In general, we believe that it is
of interest of the biologist to be able to distinguish between fusions
involving more or less expressed transcripts, because this may be
correlated with protein expression analysis.

The number of encompassing and spanning reads in s_4 sample
is comparable for both the tools, whereas for the s_7 sample
deFuse presents an higher coverage. This is mainly due to the
different alignment programs (Bowtie for deFuse and TopHat for
Bellerophontes) and to the usage of Cufflinks transcript annotations
in the proposed pipeline. In particular, our alignment and annotation
methodology reduces the probability of multiple matches, which
may lead to candidates with higher read coverage but not really
expressed.

On the other side, considering results in Table 4 in case of
fusions ANKHD1-C5orf52 (sample 018260), C1orf61-CCT3 and
MIXL1-PARP1 (sample 018266), we observe a larger number of
spanning reads supporting the validated junction. The number of
spanning reads strongly depends on the splicing detection algorithm.
Bellerophontes integrates TopHat in its pipeline, which allows to
improve the quality of spanning reads detection. This becomes
evident especially in case of junctions where deFuse shows a poor
spanning reads coverage. For instance, in the case of sample s_8
(see Table 3 of CML dataset and sample 018260 (see Table 4), the
number of Bellerophontes detected spanning is considerably higher
compared to deFuse. This result is coherent with the experimental
validation (the fusion in the 018260 sample is validated with
RT–PCR [Berger et al, 2011]).

4 DISCUSSION

The proposed framework implements a new analysis pipeline that
exploits effective alignment and annotation algorithms as well as
a filtering stage based on an accurate modeling of the junction. RNA-
Seq data analysis presents major challenges concerning ambiguous
assignments of reads to isoforms that can impact fusion detection.
Indeed, fusion candidates are discovered by means of encompassing
reads. As such, the accuracy in detection of discordant mates is a key
feature of a fusion detection pipeline. Bellerophontes introduces a
new approach where reads are mapped on experimental determined
transcripts rather than using a reference genome, thus reducing the
probability of multiple matches that may affect state-of-art fusion
detection tools based on basic alignment algorithms.

A similar effect can be noted on the analysis of spanning
reads supporting the junctions. By taking splicing events into
consideration, Bellerophontes obtain a more robust detection of the
exact junction sequence, which becomes evident especially where
the number of supporting reads is poor. Note that the number of
encompassing and spanning reads represent a discriminatory factor
to conduct a successive experimental validation of the data provided
by the software analysis.

The improved accuracy revealed another positive effect, that is the
final number of detected fusions, even in the most covered libraries
(s_7 and s_8), is compatible with a cost sensitive experimental
validation. Moreover, in s_8 three fusions involving non-annotated
genes are detected.

To improve selectivity, the pipeline integrates a set of filters,
embedding a new and more effective junction model. Indeed, this
filter discards a large number of candidates. Finally, results about
spanning read filtering and spanning-encompassing coherence check
highlight that a junction filtering is a critical step to provide an
accurate set of junctions.

Note that in all the CML samples Bellerophontes was able to
detect the expected and validated BCR-ABL1 fusion even in a
selective filtering context.

Comparative results obtained on both CML and NCBI dataset
show that Bellerophontes is able to accurately detect gene fusions
and improves the accuracy of spanning reads with respect to deFuse.
This cross-benchmarking was performed since deFuse presented the
most complete biological validation and it has shown to improve
state-of-art tools like FusionMap [Shoher et al, 2011]. The results
indicate that deFuse is not able to detect chimeric transcripts
involving non-annotated genes present in s_8. On the other side,
Bellerophontes is much more selective, reducing the total number
of fusion candidates with respect to deFuse (see Table 4).

Bellerophontes is based on pair-end reads. Other recently
proposed approaches such as FusionMap [Ge et al, 2011]
exploit long single-end reads spanning the junction. Compared to
FusionMap, our framework includes the support for new transcripts
and the accurate mapping in presence of alternative splicing events.

<table>
<thead>
<tr>
<th>Library</th>
<th>5’ Gene</th>
<th>3’ Gene</th>
<th>deFuse</th>
<th>Bellerophontes</th>
<th>deFuse</th>
<th>Bellerophontes</th>
<th>deFuse</th>
<th>Bellerophontes</th>
<th>deFuse</th>
<th>Bellerophontes</th>
</tr>
</thead>
<tbody>
<tr>
<td>018259</td>
<td>KCTD2</td>
<td>ARHGFE12</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018260</td>
<td>JMT2B</td>
<td>RHI</td>
<td>19</td>
<td>17</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018260</td>
<td>ANKHD1</td>
<td>C5orf32</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018261</td>
<td>GCLC1L1</td>
<td>PLA2G1B</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018265</td>
<td>WDR72</td>
<td>SCAMP2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018266</td>
<td>C1orf61</td>
<td>CCT1</td>
<td>54</td>
<td>37</td>
<td>17</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018266</td>
<td>MIXL1</td>
<td>PARP1</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018266</td>
<td>C1orf67</td>
<td>SLC12A7</td>
<td>43</td>
<td>40</td>
<td>24</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018266</td>
<td>GNA12</td>
<td>SHANK2</td>
<td>29</td>
<td>23</td>
<td>9</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018267</td>
<td>TLX1</td>
<td>Chor727</td>
<td>3</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018267</td>
<td>ALX3</td>
<td>RECK</td>
<td>4</td>
<td>21</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018269</td>
<td>ABL1</td>
<td>BCR</td>
<td>91</td>
<td>89</td>
<td>14</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>018269</td>
<td>NUP214</td>
<td>XRRA</td>
<td>67</td>
<td>58</td>
<td>15</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCIH660</td>
<td>TMPRSS2</td>
<td>ERG</td>
<td>19</td>
<td>23</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*All the library identifiers, with exception of the last row, refer to the accession number
reporting the SRR prefix in the NCBI databank.
We also performed a comparative analysis of Bellerophontes against deFuse, TopHat-Fusion (http://tophat-fusion.sourceforge.net) and Chimerascan (Ver et al. 2011), on the CML samples. All of them recognized BCR-ABL1 and provided a set of additional fusions. Interestingly, there was almost no overlap among all the tools in respect of these additional fusions (Fig. 10B). For this reason, most of them are likely to be false-positive. It can be noted that Bellerophontes has the smaller set of this additional fusions. Finally, we in order to further evaluate the performance of the proposed method we performed analysis on an additional dataset, that is MCF7 cell line by Edgren et al. (2010), for which we found all validated fusions, namely, BCAS4-BCAS3, ARFGEF2-SULF2 and RPS6KB1-TMEM49.

5 SOFTWARE
Bellerophontes is implemented in Java/Perl/Bash language. It runs on a standard Linux machine and it fully supports multithreaded computation. The performance strongly depends on the memory and number of CPUs available. However, the analysis on the proposed datasets have been performed with an Intel I7 TM920 at 2.67 GHz and 16 GB of RAM Memory.

Conflict of Interest: none declared.

REFERENCES