FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution

Huanying Ge1,*, Kejun Liu2, Todd Juan3, Fang Fang4, Matthew Newman2 and Wolfgang Höck1

1Research and Development Informatics, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, 2OmicSoft Corp., 2500 Gateway Centre Blvd, Suite 550B, Morrisville, NC 27560, 3Protein Science, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320 and 4Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

ABSTRACT

Motivation: Next generation sequencing technology generates high-throughput data, which allows us to detect fusion genes at both transcript and genomic levels. To detect fusion genes, the current bioinformatics tools heavily rely on paired-end approaches and overlook the importance of reads that span fusion junctions. Thus, there is a need to develop an efficient aligner to detect fusion events by accurate mapping of these junction-spanning single reads, particularly when the read gets longer with the improvement in sequencing technology.

Results: We present a novel method, FusionMap, which aligns fusion reads directly to the genome without prior knowledge of potential fusion regions. FusionMap can detect fusion events in both single- and paired-end datasets from either RNA-Seq or gDNA-Seq studies and characterize fusion junctions at base-pair resolution. We showed that FusionMap achieved high sensitivity and specificity in fusion detection on two simulated RNA-Seq datasets, which contained 75 nt paired-end reads. FusionMap achieved substantially higher sensitivity and specificity than the paired-end approach when the inner distance between read pairs was small. Using FusionMap to characterize fusion genes in K562 chronic myeloid leukemia cell line, we further demonstrated its accuracy in fusion detection in both single-end RNA-Seq and gDNA-Seq datasets. These combined results show that FusionMap provides an accurate and systematic solution to detecting fusion events through junction-spanning reads.

Availability: FusionMap includes reference indexing, read filtering, fusion alignment and reporting in one package. The software is free for noncommercial use at (http://www.omicsoft.com/FusionMap).

Contact: gei@amgen.com

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 1, 2011; revised on May 5, 2011; accepted on May 11, 2011

Introduction

A fusion gene is a hybrid gene by joining parts from two previously separate genes at transcript or genomic level. The importance of fusion genes in cancer development has been well recognized since the discovery of the recurrent BCR-ABL1 fusion gene in chronic myelogenous leukemia (CML) (TKachuk et al., 1990) and TMPRSS2-ERG fusion gene in solid tumor (Tomlins et al., 2005). Fusion genes are caused by chromosomal aberrations, e.g. inversion, translocation, large deletion or insertion. For example, BCR-ABL1 is formed by a translocation event involving chromosome 9 and 22. Its fusion transcript is translated into an abnormal tyrosine kinase, which plays a critical role in the development of CML (Lugo et al., 1990; Rabbitts, 2009).

These chromosomal aberrations were traditionally detected through fluorescence in situ hybridization (FISH) or comparative genomic hybridization (CGH) techniques (Edwards, 2009). Array-CGH and SNP array were later developed to detect genomic copy number variations at 1Kb resolution in a high-throughput fashion (Pinkel et al., 1998; Redon et al., 2006). Compared to these techniques, next-generation sequencing (NGS) generates base-pair resolution data, which allows the detection and characterization of genomic aberrations at more levels of details. With NGS, fusion genes were detected using either genomic DNA sequencing (gDNA-Seq) (Campbell et al., 2008; Hampton et al., 2009) or transcriptome sequencing (RNA-Seq) (Berger et al., 2010; Levin et al., 2009; Maher et al., 2009a).

In NGS datasets, fusion genes can be detected based on both paired- and single-end reads. On the one hand, paired-end reads (50–100 bp) generated from long fragments (200–500 bp) are shown to be useful in increasing the ability to detect fusion events (Berger et al., 2010). In a paired-end NGS dataset, a discordant read pair is one that is not aligned to the reference genome with the expected distance or orientation. If a set of discordant read pairs are mapped to two different genes, a fusion gene is suggested (Maher et al., 2009b). Computational tools like FusionSeq (Shomer et al., 2010) have been developed to detect fusion candidates by analyzing paired-end reads and removing spurious candidates using varied filters. On the other hand, single-end reads that span the fusion junctions provide base-pair evidence for the fusion events. Since sequencing technology is improving rapidly, the current usable read length from the Illumina Genome Analyzer is typically in the range of 75–100 nt and reaches 150 nt using Illumina’s TruSeq SBS V3 GA Kit (Illumina, 2010). Long read lengths not only increase the number of fusion junction-spanning reads but also allow us to align them unambiguously to the fusion junction. Moreover, junction-spanning reads can also be present in paired-end datasets. They provide additional information

1 INTRODUCTION
for fusion genes suggested by discordant read pairs. When the read length is long or the inner distance (the gap size) between read pairs is short, the number of junction-spanning reads would be larger than that of discordant read pairs. Then, fusion detection based on junction-spanning reads is more powerful than the paired-end approach.

One approach to detect fusion junction-spanning reads is mapping the reads to a set of artificially constructed exon-exon segments from potential fusion gene pairs (Levin et al., 2009). However, such an approach relies on prior knowledge of potential fusion regions and only identifies fusion junctions between known exons.

Another approach is using splice aligners, such as TopHat (Trapnell et al., 2009) SpliceMap (Aa et al., 2010) and MapSplice (Wang et al., 2010), to detect splice junctions. These alignment tools can split and map a read to different locations in a genome but only within local regions (usually ≤50 Kb). They are not able to identify distant or intra-chromosome fusion genes. Therefore, the development of a dedicated fusion aligner to align junction-spanning reads is necessary for fusion detection.

In this article, we describe a new method, FusionMap, which detects and aligns fusion junction-spanning reads to the reference genome systematically. FusionMap can be applied to both single- and paired-end datasets from either RNA-Seq or gDNA-Seq studies. During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

2 METHODS

2.1 Basic alignment

Fusion reads are defined as reads spanning fusion junctions between two different genes or genomic regions. FusionMap makes substantial use of a basic aligner to detect and align fusion reads to the genome reference. The basic alignment is an implementation of a modified GSNAP method (Wu and Nacu, 2010). GSNAP is a hash table based aligner allowing for nonunique mapping and reads spanning fusion junctions between genes or genomic regions.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.

During the fusion alignment, FusionMap dynamically creates a pseudo fusion transcript/sequence library based on consensus junction sequences suggested by seed reads, and then aligns the remaining full-length reads to this pseudo reference (Fig. 1 and Supplementary Fig. S1). The program reports a list of detected fusion junctions, statistics of supporting reads, fusion gene pairs, as well as genomic locations of breakpoints and junction sequences, which characterize fusion genes comprehensively at base-pair resolution.
Table 1. Important parameters in fusion alignment

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum length of a seed read $a$</td>
<td>25</td>
</tr>
<tr>
<td>Maximum hits of a seed read $β$</td>
<td>1 (unique mapping)</td>
</tr>
<tr>
<td>Non-canonical splice pattern penalty (G)</td>
<td>2 (equal to two mismatches)</td>
</tr>
<tr>
<td>Minimum distance of a fusion junction</td>
<td>5000 bp</td>
</tr>
<tr>
<td>Maximum alignment penalty for a read</td>
<td>max(2, (ReadLength−31)/15)</td>
</tr>
<tr>
<td>Minimum number of rescued reads</td>
<td>1</td>
</tr>
<tr>
<td>Minimum number of distinct fusion reads</td>
<td>2</td>
</tr>
</tbody>
</table>

alignment penalty for $S$ suffers the total number of mismatches plus the gap penalty (default gap penalty is 2), and define $F_0(S)$ as the maximum allowable alignment penalty. By default, $F_0(S) = \max(2, (\text{ReadLength}−31)/15)$.

In the RNA-Seq datasets, studies have shown that most of detected fusion junctions in RNA-Seq datasets are splices junctions of fusion genes (Levin et al., 2009; Shomer et al., 2010). Thus, we also define an additional penalty, $G(i)$, for the mapped genomic location of the cutting position $i$ in RNA-Seq alignment:

$$G(i) = \begin{cases} 0, & \text{if alignment of } S_{i−2} \text{ and } S_{i+1}, \text{ or } S_{i+1}, \text{ or } \Omega \neq \emptyset, \\ 1, & \text{if alignment of } S_{i−2} \text{ and } S_{i+1}, \text{ or } \Omega = \emptyset. \end{cases}$$

where $\Omega$ denotes the event that mapped locations of $S_{i−2}$ and $S_{i+1}$, or $S_{i+1}$, have one of canonical splice patterns (GT-AG, GC-AG and AT-AC); $G$ is a user-defined parameter, representing the noncanonical splice pattern penalty.

Based on the alignment of $S_{i−2}$ and $S_{i+1}$, FusionMap determines the optimal cut($i^∗$) to the fusion position(s) for $S_{i−2}$ if all of the following conditions are true:

$$r^∗ = \arg\min_{r<25} \left\{ F(S_{i−1}) + \Omega(S_{i+1}, \Omega) + G(i) \right\},$$

where $\Omega(S_{i+1}, \Omega) = \left\{ \begin{array}{ll} 0, & \text{if alignment of } S_{i−2} \text{ and } S_{i+1}, \text{ or } \Omega \neq \emptyset, \\ 1, & \text{if alignment of } S_{i−2} \text{ and } S_{i+1}, \text{ or } \Omega = \emptyset. \end{array} \right.$

FusionMap excludes any candidate seed read if no $i$ satisfies these conditions. By default, $G = 2$ for RNA-Seq datasets. FusionMap favors the cut and alignment that forms splice junctions by setting $G > 0$. When we set $G > F_0(S_{i−2}, S_{i+1})$, it only detects fusion junctions with canonical splice patterns. For optimal fusion cut($i^∗$), we also require that the number of alignment hits from each single end ($S_{i−2}$ and $S_{i+1}$) is not greater than a user-defined parameter $β$ (the maximum hits of a read ends to control the specificity. By default, we set $β = 1$ which requires unique alignment of each end, controlling the false discovery of fusion events in the repetitive regions. The alignment of $S_{i−2}$ and $S_{i+1}$ may have different strand orientations and this information is important to define a fusion. In this step, qualified reads are confirmed as seed reads.

The number of seed reads provides confident estimate for the abundance of a fusion gene or transcript. Denote $M_{\text{mapped}}$ as the total number of reads in the NGS dataset and $M_{\text{unmapped}}$ the number of reads mapped to the reference before fusion alignment. In the remaining unmapped reads, suppose $M_{\text{unmapped}}$ is the number of seed reads detected by FusionMap for a particular fusion junction. We define a measure called seed reads per kilobase of seed region and per million mapped reads (SRPKM):

$$\text{SRPKM} = \frac{M_{\text{mapped}}}{L \times M_{\text{unmapped}}} \times 10^3 \times 10^8,$$

where $L$ is the seed region length adjusted by the number of allowable mismatches. For 75 nt reads, $L = 75 - 2 \times 25 + 1 = 2^2 + 30$ when using $a = 25$ and allowing two mismatches; SRPKM = 1 when FusionMap detects one seed read for the fusion junction in a dataset with 33 million mappable reads.

3 RESULTS

3.1 Evaluation of FusionMap on simulated datasets

We evaluated the performance of FusionMap on paired-end RNA-Seq datasets from simulation. The simulation used H1 human embryonic stem cells (hESCs) as backgrounds, which were not expected to harbor any fusion transcripts. Two background datasets (2 × 75 nt) were downloaded from NCBI Sequence Read Archive (SRA) under accession no. SRR065491 and SRR066679, which were generated by the ENCODE Caltech RNA-Seq project (Birney et al., 2007; Raney et al., 2010). We simulated two datasets P1 and P2 based on SRR065491 and SRR066679, respectively (Table 2). Each dataset contained 75 nt paired-end reads from 50 simulated fusion transcripts and they were mixed with the background reads. The number of reads generated per fusion depended on the RPKM values of the genes forming the fusion transcript, and their insert fragment size was equal to the median insert size in the corresponding background dataset.

We aligned both datasets to human genome assembly hg19 and a splice junction library using Bowtie (version 0.12.5) (Lang et al., 2009) allowing two mismatches. The splice junction library was generated based on RefSeq gene models. After this regular
FusionMap

Table 2. Simulated RNA-Seq datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Background (H1-hESC 2 × 75 nt)</th>
<th>Median insert size</th>
<th>#Background reads</th>
<th>#Simulated reads</th>
<th>#Fusion reads</th>
<th>#Total Reads</th>
<th>Alignment statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>#Mappable reads</td>
</tr>
<tr>
<td>P1</td>
<td>SRR065491 400 bp</td>
<td>78652192</td>
<td>179886</td>
<td>70414</td>
<td>78831278</td>
<td>38195304</td>
<td>25991440</td>
</tr>
<tr>
<td>P2</td>
<td>SRR066679 158 bp</td>
<td>49179164</td>
<td>114418</td>
<td>8600</td>
<td>49293582</td>
<td>32193918</td>
<td>12599293</td>
</tr>
</tbody>
</table>

Fusion reads are reads generated from fragments covering fusion junction.

alignment, unmapped reads were supplied as the input to run FusionMap, using the following parameter combinations: the minimum end length of a seed read $\alpha = 20, 25$ or $30$, the maximum hits of a read end $\beta = 1$, $2$ or $5$ and the penalty for noncanonical splice pattern $G = 2$ or $4$.

To assess the performance of FusionMap, we counted the numbers of false/true positives and calculated sensitivity/specificity at varied cutoffs on the number of seed reads (Supplementary Materials). In both datasets, FusionMap achieved high sensitivity and specificity in fusion detection using the parameters $\{\alpha = 25, \beta = 1\}$ (Fig. 2A and 2B). $\alpha$ and $\beta$ are two important parameters in determining FusionMap’s sensitivity and specificity. At the same cutoff on the number of seed reads, the specificity was reduced and the sensitivity was increased when we decreased the minimum end length of the seed read ($\alpha$) or allowed more alignment hits for each read end ($\beta$) (Supplementary Fig. S5). We also assessed the performance of FusionMap without the advanced filtering step, and the results showed that the specificity decreased a little. It demonstrated that the great majority of high specificity was achieved through the fusion alignment of junction-spanning reads.

We next evaluated the impact of parameter $G$ on the fusion detection accuracy. Because the maximum alignment penalty is 2 (number of mismatches plus the gap penalty) for a 75 nt read, fusion junctions with canonical splice patterns were favored by setting $G = 2$ and became the only choice for FusionMap detection by setting $G = 4$. Thus the specificity was improved in dataset P1 and P2 by increasing $G$ from 2 to 4 (Fig. 2A and 2B). Moreover, since it is possible that the fusion junction may not have the canonical splice patterns, we used the same simulation scheme to get an additional dataset (named P3) with breakpoints in the middle of exons (see Supplementary Materials). We found that the majority of simulated fusion genes (42 out of 50) were detected using $\{\alpha = 25, \beta = 1, G = 2\}$, and the sensitivity was slightly impaired using $G = 2$ comparing to that from $G = 0$ (Supplementary Fig. S6A).

We assessed the accuracy of fusion breakpoint detection by calculating the distance between the genomic coordinates of detected and simulated breakpoints. Among breakpoints detected for simulated fusion genes using $G = 2$, close to 70% matched the exact positions in dataset P1 and P2 (Fig. 2C and 2D). We further examined the breakpoints with the genomic distance larger than five base-pairs and found the majority of them were actually close ($<5\text{ bp}$) to the simulated positions at the transcript levels. The detected and simulated breakpoints were in two adjacent exons but far away at the genomic level, which caused the aggregation step to fail to combine them into one. In fact, for the fusion gene with multiple detected junctions, if we only chose the one with the maximum number of seed reads, $>96\%$ of them matched the exact positions in dataset P1 and P2. Moreover, when using $G = 4$, FusionMap detected one fusion junction per fusion gene, and the detection accuracy reached 100% (the red bar in Fig. 2C and 2D).
We compared FusionMap with FusionSeq (ver. 0.6.1) (Sboner et al., 2010) on simulated datasets P1 and P2. FusionSeq contains a detection module to identify fusion candidates based on inter-transcript paired-end reads, a filtration cascade module to remove spurious candidates and a junction-sequence identifier to detect the exact sequence at breakpoints. We applied the detection module to uniquely aligned read pairs from the Bowtie alignment outputs. In the detection step, we set the cutoff for the minimum number of inter-transcript paired-end reads to be either 2 or 4. In the filtering steps, we noticed that the abnormal insert size filter had a major effect on the result if the inner distance between read pairs was large, e.g. in dataset P1. Thus, we ran three filtering workflows with abnormal insert size P-value cutoff of 10^{-3}, 10^{-6} or without the insert size filter (See workflow scripts in Supplementary Materials).

To make a fair comparison, we further cleaned the fusion reports from FusionSeq using FusionMap’s advanced filter. Then the numbers of false/true positives were counted at varied cutoffs on the DASPER score (the difference between the observed and analytically calculated expected supportive paired-end reads) which was generated by FusionSeq. As shown in Figure 3A and 3B, FusionMap achieved better sensitivity and specificity in both datasets, particularly in dataset P2. Because each read pair was generated from a long-sequence fragment (∼158 bp in P1 and ∼400 bp in P2), a fusion event was detected through junction-spanning read in FusionMap if the junction was in the middle of one read, and was detected through inter-transcript paired-end reads in FusionSeq if the junction was in the inner part between the two reads. In Figure 3C and 3D, we illustrated the number of seed reads detected by FusionMap using α = 25 and that of inter-transcript read pairs detected by FusionSeq, separately, against the number of simulated fragment covering the fusion junction. Based on the regression coefficients, the ratio between the number of seed reads and inter-transcript read pairs were close to the theoretical values: (75 − 25 × 2 × 2/(400 − 75 × 2)) = 0.2 in dataset P1 and (75 − 25 × 2 × 2/(158 − 75 × 2)) ≈ 6 in dataset P2. FusionMap achieved substantially higher sensitivity than paired-end approach when the inner distance was small.

Both FusionMap and FusionSeq were run on a 64-bit machine with Intel Xeon E5640 (2.66 GHz, two Quad Cores) and 12 GB RAM. As shown in Table 3, FusionMap ran faster than the FusionSeq workflow but required more memory to cache the human genome reference during the alignment.

### 3.3 Fusion gene discovery in K562 cell line

We further applied FusionMap to two published NGS datasets (named A and B) that sequenced the K562 CML cell line. Dataset A contains 14 million 76 nt single-end reads sequencing the captured cDNA fragments that were specific to 467 cancer-related genes (Levin et al., 2009). Dataset B contains 160,886 pairs of 38 nt reads from a targeted gDNA-Seq study which used M-Bcr [the
A, we identified 1003 junction-spanning reads connecting the end of exon BC and fusion junction-spanning reads (black) linking gene BC in dataset B. (Fig. 4B and C). The 5 bp junction window was caused by the same 'GAGTG' sequence in the region connecting two genes. The genomic fusion breakpoint was in the intron region right after the 14th exon in BC and in the intron region between exons 1a and 1b in ABL1. The result indicated that exon 1b of ABL1 had been skipped during the transcription of BCR-ABL1 fusion gene. The detected fusion junctions were in complete agreement with the breakpoints reported previously (Levin et al., 2009; Shibata et al., 2010).

In dataset A, we also detected 164 junction-spanning reads connecting the end of exon 29 in NUP214 to the beginning of exon 2 in XKR3 and two additional junctions: NUP214 (exon 29)-XKR3 (exon 5) and NUP214 (exon 29)-XKR3 (exon 4) with 4 and 9 supporting reads. One more fusion junction, NUP214 (exon 27)-XKR3 (exon 2), can be detected with three supporting fusion reads if we used $\alpha = 15$ instead of 25 and relaxed filters on fusion read numbers. These detected fusion junctions are in agreement with previous results (Levin et al., 2009). The detected junction positions of NUP214-XKR3 were all on known exon boundaries, indicating that these fusion transcripts were alternative splicing isoforms of the NUP214-XKR3 fusion gene.

Moreover, in dataset A, FusionMap also detected SNHG3/RCC1-PICALM, PRIM1-NACA, NCKIPSD-CELSR3 and SLC29A1-HSP90AB1 fusion genes (Supplementary Table S1). The inner distances between two breakpoints on the genome are ∼19 Kb in PRIM1-NACA, ∼21 Kb in NCKIPSD-CELSR3 and ∼15 Kb in SLC29A1-HSP90AB1. They are close neighbors on the genome and are likely to be fusion transcripts caused by read-through events.

4 DISCUSSION

The FusionMap method described in this article allows us to align fusion junction-spanning reads to references and can, therefore, fully utilize NGS data to characterize fusion genes at base-pair resolution. The method achieves the following three goals in computational detection of fusion genes using NGS data.

First, FusionMap performs a sensitive and complete search of fusion junction-spanning reads without relying on any additional information. It has the same detection power in both single- and paired-end NGS datasets. In paired-end datasets, the fusion report from junction-spanning reads is complementary to that from discordant read pairs. Moreover, as shown by dataset P2, fusion detection based on junction-spanning reads is more powerful than the paired-end approach when the inner distance between read pairs is short. In theory, for a set of 75 nt paired-end reads generated from a 400 bp fragment library, the ratio between the number of junction-spanning reads and discordant read pairs is close to $75^2/1000 = 0.6$. The ratio will increase when the usable read length increase with new sequencing techniques or reagents (Ilumina, 2010), making FusionMap more useful and effective.

Second, FusionMap can be applied to both gDNA-Seq and RNA-Seq datasets. Compared to gDNA-Seq datasets, RNA-Seq is more complicated to detect fusion genes due to the implication of exon–exon junctions. In gDNA-Seq fusion junctions detected are the actual fusion positions in the genome, while junctions detected in RNA-Seq are more likely to be splice junctions rather than genomic
fusion positions. Therefore, FusionMap increases the specificity by adopting the noncanonical splice pattern penalty \( G \) for RNA-Seq datasets. SRPKM is a normalized measure across different read lengths and different samples. Its calculation mimics the RPKM calculation for each gene. In RNA-Seq datasets, the SRPKM for each fusion junction provides the relative abundance of fusion transcript isoforms.

Third, FusionMap reports the exact junction sequences, which characterize fusion genes at base-pair resolution. Based on the junction sequence, researchers can design primers to confirm the fusion transcript by RT-PCR or the fusion breakpoint on the genome using genomic PCR. Furthermore, detected junction sequences in the gDNA-Seq dataset might be useful to investigate the mechanism of homologous recombination that creates the fusion product, like the 5 bp junction window of the ‘GAGTG’ sequence in BCR-ABL1 in dataset B.

Among the advantages of our method, FusionMap is independent from any mapping method used for regular alignment. It can be applied directly to the whole dataset but would take less time if it uses unmapped reads after the regular alignment. In this article, we showed the performance of FusionMap on unmapped reads after the Bowtie alignment. We also tried FusionMap based on TopHat (version 1.0.14) (Trapnell et al., 2009) outputs, and got almost the same results in datasets P1 and P2. However, in dataset A, if we used TopHat for regular alignment, the three read-through fusion genes were missing in FusionMap’s result because they were close neighbors on the genome (<50 Kb) and were identified as novel splice junctions by TopHat.

Moreover, FusionMap is not limited to a single dataset, but rather can be applied to multiple NGS datasets in a single run. It provides an easy way to detect recurring fusion events by grouping fusion junctions by rows in one table.

The main limitation of FusionMap is its reliance on read length. FusionMap successfully detects the true genomic fusion in dataset B (gDNA-Seq dataset with 38 nt reads for K562 CML) with high specificity, mainly because it is a targeted sequencing dataset for BCR-ABL1 only. In order to achieve high sensitivity and specificity, read lengths \( \geq 57 \) nt are preferred in whole genome/transcriptome sequencing datasets. FusionMap will benefit from longer read length and deeper coverage.

We also noticed that the current version of FusionMap requires longer computing time than existing splice aligners. It is our plan to improve the search strategy for fusion junctions by adopting the current detection methods for splice junctions (Au et al., 2010; Wang et al., 2010), extending base by base following a hit of a potential seed read. The modification will speed up the fusion alignment step.

5 SOFTWARE

FusionMap is implemented in C\# and runs under Windows (or Linux using MONO). It can run in both 32- and 64-bit modes. The 64-bit mode performs much faster but it requires at least 6 GB RAM.

ACKNOWLEDGEMENTS

This work benefited from discussions with Kim Quon in Amgen Inc. We thank Keith Weinstock and Robert DuBose for proofreading the manuscript. We thank Barbara Wold and UCSC ENCODE team for making their data available for our simulation study. We also thank the anonymous reviewers for their valuable suggestions.

Conflict of interest: K.L. and M.N. have financial interest in OmicSoft Corporation. The research reported in this paper is in the field in which OmicSoft is developing commercial products.

REFERENCES


