BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1922–1928
doi:10.1093/bioinformatics/btr310

Sequence analysis Advance Access publication May 18, 2011

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

Huanying Ge[1,*], Kejun Liu[2], Todd Juan[3], Fang Fang[4], Matthew Newman[2] and Wolfgang Hoeck[1]

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

Associate Editor: Alex Bateman

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: ge@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

1 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 (Sboner *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 V5 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

*To whom correspondence should be addressed.

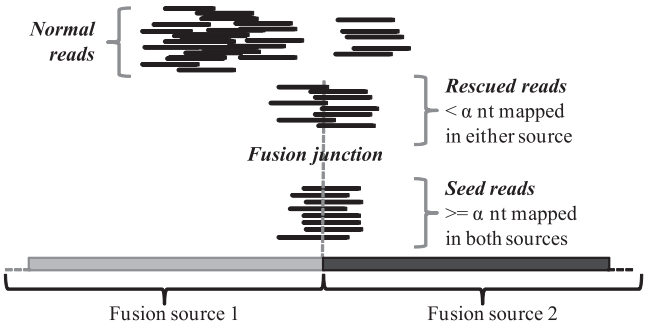

Fig. 1. Identification of fusion events in NGS dataset through junction-spanning reads. Fusion source 1 and 2 are from two different genes or genomic regions.

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 (Au *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 $\leq$50 Kb). They are not able to identify distant or inter-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 multiple mismatches and long indels. In GSNAP, the hash table indexes 12mers every 3 nt in the genome and uses another array to store the position list of each 12mer. Each 12mer is used as a hash key to quickly look up the possible positions for the input read. GSNAP can detect complex variation such as long indels and novel splicing junctions in both gDNA-Seq and RNA-Seq datasets. In FusionMap, we have implemented GSNAP with two important modifications. First, we store the positions for 14mers instead of 12mers in the hash table. The memory required for the data structures increases from 5.0 GB to 5.9 GB, but the position list for each 14mer is decreased by a factor of 16 on average. This modification speeds up the GSNAP alignment by a factor of 2–4 times depending on the read length. Second, for RNA-Seq alignment where a known gene model is available, we first align the reads to the transcriptome library allowing for nonunique mapping, and remap the transcriptome coordinates to genomic coordinates. Most of the nonunique transcriptome coordinates will be converted to unique genome coordinates. This modification significantly improves the performance of RNA-Seq alignment, as the majority of the reads can be mapped to the transcriptome as exon reads or known exon junction spanning reads. For those reads that cannot be mapped to the transcriptome, we use GSNAP to align them to the genome as potential intron or novel inter-exon reads.

2.2 Fusion detection pipeline

FusionMap can be applied directly to the whole dataset but would take less time if it focuses on unmapped reads after regular alignment. The fusion detection pipeline contains five main steps as described in the following subsections:

Step 1: *filtering*: FusionMap filters input reads before fusion alignment. For each read, if the quality score is available, FusionMap has an option to trim the low-quality base from the 3′ end until the first high-quality base is found. It then aligns all reads to a pre-built reference index. The reference can be a genome or a set of target regions. The alignment considers known and novel splice junctions if the input reads are from RNA-Seq studies. All aligned reads will be regarded as normal reads from transcripts or genomic regions and filtered out. Due to sequencing/base-calling errors and single nucleotide polymorphisms (SNPs), some normal reads will not be filtered out using standard alignment parameters, e.g. allowing two mismatches for 75 nt reads. During fusion alignment as described in Step 2, FusionMap will cut each unfiltered normal read into two parts and align them separately. The program may detect a false fusion junction if one part is aligned to its true position while the other part is misaligned to a different location due to the sequence similarity. In order to reduce the false discovery due to the misalignment, by default, we relax the alignment parameter in this filtering step allowing at most 8% dissimilarity between each read and the reference. It allows at most six mismatches for a 75 nt read.

Moreover, because a true fusion read spans across a fusion junction, if we cut the read in the middle we should be able to align at least one half to the reference. If neither half can be aligned, the read will be regarded as a low-quality read or artifact and filtered out. In practice, more than 90% of the input reads can be excluded from further consideration by the two filtering strategies.

Step 2: *fusion alignment of seed reads*: FusionMap detects fusion junctions based on seed reads which contain the fusion positions in the middle region of the reads (Fig. 1). In order to search seed reads efficiently, FusionMap first follows a simple rule to check each unmapped read. Given a read string $\mathbb{S}_{1\sim N}$ with N nucleotides, let $\mathbb{S}_{i\sim j}$ denote the substring of the whole read from the i-th to j-th nucleotide and let parameter α denote the minimum end length of a seed read. If the α-nt prefix ($\mathbb{S}_{1\sim\alpha}$) and suffix ($\mathbb{S}_{(N-\alpha+1)\sim N}$) of a read can both be aligned to the reference, FusionMap then considers it as a candidate seed read.

For each candidate seed read, FusionMap cuts the whole read into two ends: $\mathbb{S}_{1\sim i}$ and $\mathbb{S}_{(i+1)\sim N}$, where $\alpha \leq i \leq$ N-α. The region [α, N-α] contains all possible cutting positions, and is called the seed region. By default, $\alpha = 25$, a 75 nt read has 26 possible cutting positions while a 100 nt read has 51 possible cutting positions. After cutting, FusionMap aligns $\mathbb{S}_{1\sim i}$ and $\mathbb{S}_{(i+1)\sim N}$ to the reference separately. Let $\mathbb{S}$ denote the general term for a short sequence, either a whole read or a substring of a read. We then define F($\mathbb{S}$) as the

Table 1. Important parameters in fusion alignment

Parameter	Default value
Minimum end length of a seed read α	25
Maximum hits of a seed read end (β)	1 (unique mapping)
Non-canonical splice pattern penalty (G)	2 (equal to two mismatches)
Minimum distance of a fusion junction	5000 bp
Maximum alignment penalty for a read	max(2, (ReadLength-31)/15)
Minimum number of rescued reads	1
Minimum number of distinct fusion reads	2

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

In the RNA-Seq datasets, studies have shown that most of detected fusion junctions in RNA-Seq datasets are splice junctions of fusion genes (Levin *et al*., 2009; Sboner *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:

$$G(i) = \begin{cases} G, \text{if alignment of } \mathbb{S}_{1\sim i} \text{ and } \mathbb{S}_{(i+1)\sim N} \notin \Omega \\ 0, \text{if alignment of } \mathbb{S}_{1\sim i} \text{ and } \mathbb{S}_{(i+1)\sim N} \in \Omega \end{cases},$$

where Ω denotes the event that mapped locations of $\mathbb{S}_{1\sim i}$ and $\mathbb{S}_{(i+1)\sim N}$ 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 $\mathbb{S}_{1\sim i}$ and $\mathbb{S}_{(i+1)\sim N}$, FusionMap determines the optimal cut(s) (i^*) as the fusion position(s) for $\mathbb{S}_{1\sim N}$ if all of the following conditions are true:

$$\begin{cases} i^* = \arg \min\limits_{\alpha \le i \le N-\alpha} \left\{ F(\mathbb{S}_{1\sim i}) + F(\mathbb{S}_{(i+1)\sim N}) + G(i) \right\} \\ F(\mathbb{S}_{1\sim i^*}) + F(\mathbb{S}_{(i^*+1)\sim N}) + G(i^*) \le F_0(\mathbb{S}_{1\sim N}) \\ F(\mathbb{S}_{1\sim i^*}) \le F_0(\mathbb{S}_{1\sim i^*}) \\ F(\mathbb{S}_{(i^*+1)\sim N}) \le F(\mathbb{S}_{(i^*+1)\sim N}) \end{cases}$$

FusionMap excludes any candidate seed read if no i satisfying 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 $G > F_0(\mathbb{S}_{1\sim N})$, it only detects fusion junctions with canonical splice patterns. For optimal fusion cut(s) i^*, we also require that the number of alignment hits from each single end ($\mathbb{S}_{1\sim i^*}$ and $\mathbb{S}_{(i^*+1)\sim N}$) is not greater than a user-defined parameter β (the maximum hits of a read end) to control the specificity. By default, we set $\beta = 1$ which requires unique alignment of each end, controlling the false discovery of fusion events in the repetitive regions. The alignment of $\mathbb{S}_{1\sim i^*}$ and $\mathbb{S}_{(i^*+1)\sim N}$ 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_{total} as the total number of reads in the NGS dataset and M_{mapped} as the number of reads mapped to the reference before fusion alignment. In the remaining unmapped reads, suppose M_{seed} 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):

$$\text{SRPKM} = \frac{M_{seed}}{L \times M_{mapped}} \times 10^3 \times 10^6,$$

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

Step 3: *fusion junction searching and rescue mapping*: FusionMap searches possible fusion junction position(s) based on the consensus of mapped fusion positions from seed reads. It only considers fusion junctions with a distance larger than 5000 bp by default. FusionMap creates a pseudo fusion transcript/sequence library by extracting and extending both fusion source sequences and concatenating them into a single pseudo transcript/sequence reference. It dynamically builds the reference index based on the pseudo fusion library and aligns unmapped possible fusion reads to this pseudo reference. Reads mapped in this step are considered as rescued reads (Fig. 1) providing additional support for the detected fusion junction. By default, FusionMap only reports fusion candidates with at least one rescued fusion read (Table 1).

Step 4: *aggregation of fuzzy fusion junctions*: Because nucleotides at the edges of two fusion sources might be similar, as well as the allowance for mismatch and gap during the alignment (see one example in the Supplementary Fig. S2), FusionMap may detect multiple optimal cut positions on the seed read and then report a set of fuzzy junctions close to the true one. Usually, these junctions are only differed by losing (gaining) a few (usually 1–5) base-pairs on one fusion edge and gaining (losing) the equal amount of base-pairs on the other fusion edge. In order to reduce redundancy, FusionMap aggregates these junctions and reports the one having the most seed reads, or having the most rescued reads if their numbers of seed reads are the same. The seed and rescued reads belonging to the same fuzzy junction set are also aggregated for the final fusion report.

For each fusion junction, FusionMap counts the number of supportive reads with distinct mapped locations. By default, FusionMap only reports fusion candidates with at least two distinct fusion reads, removing potential false positives resulted from the PCR amplification of the same read.

Step 5: *advanced filtering*: Detected fusion candidates can go through this optional step to further filter out potential false positives. We have prepared a blacklist for genes and gene pairs. The current gene blacklist includes mitochondrial and ribosomal genes according to Gene Ontology (GO), and pseudogenes according to three sources: Ensembl annotations, Entrez Gene Database and HUGO Gene Nomenclature Committee (HGNC). Fusion candidates involving genes included in this blacklist will be removed. The current gene pair blacklist includes the gene family definition according to HGNC and the paralog gene pairs from Ensembl BioMart (Smedley *et al*., 2009). Fusion candidates formed by gene pairs that are paralogs or from the same gene family will be removed.

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) (Langmead *et al*., 2009) allowing two mismatches. The splice junction library was generated based on RefSeq gene models. After this regular

Table 2. Simulated RNA-Seq datasets

Dataset	Background (H1-hESC 2×75 nt)	Median insert size	#Background reads	#Simulated reads	#Fusion reads	#Total Reads	Alignment statistics		
							#Mappable reads	#Uniquely mapped and paired reads	#Unmapped Reads
P1	SRR065491	400 bp	78652192	179086	70414	78831278	52839838	38195304	25991440
P2	SRR066679	158 bp	49179164	114418	8600	49293582	36734289	32193918	12559293

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). α and β 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 (α) or allowed more alignment hits for each read end (β) (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 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).

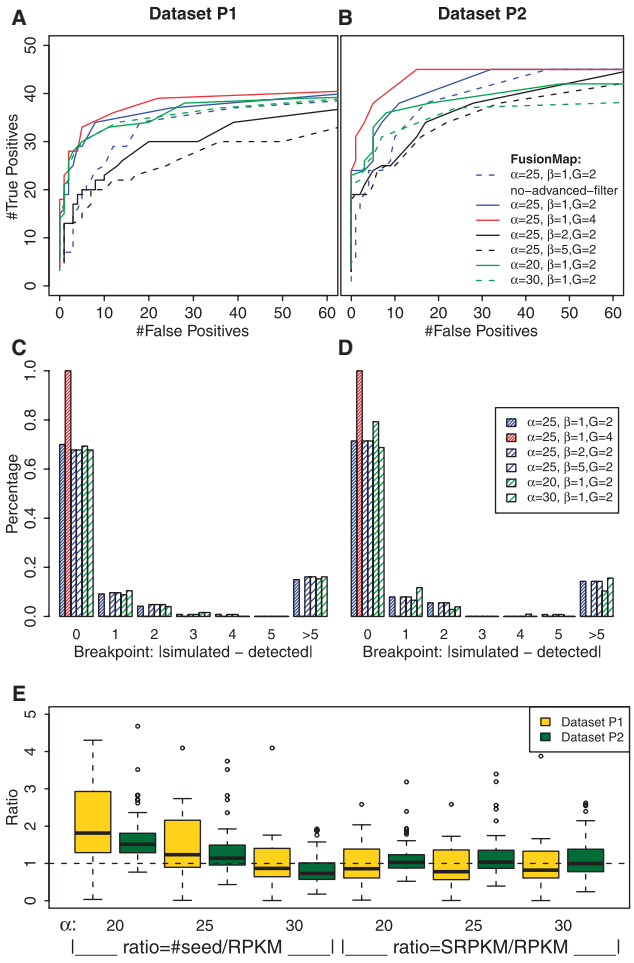

Fig. 2. Performance of FusionMap on simulated RNA-Seq datasets. (**A** and **C**): dataset P1; (**B** and **D**): dataset P2. (A and B): number of false and true positives in results using varied parameter combinations. (C and D): detection accuracy of breakpoint positions measured by the distance between simulated and detected positions. (**E**): estimation of RPKM values for simulated fusion genes using the number of seed reads or SRPKM values. The ratios were calculated for each fusion gene based on results using three α values and $\{\beta = 1, G = 2\}$ in both datasets.

We also assessed the accuracy in the simulated dataset with fusion breakpoints in the middle of exons (dataset P3). The results showed that FusionMap detected one fusion junction per fusion gene, and >70% junctions matched the simulated positions and 100% were within ±5 bp range (Supplementary Fig. S6B). In dataset P3, the breakpoint detection accuracy was slightly impaired by setting G = 2 comparing to that from G = 0 since FusionMap favored the canonical splice sites if the simulated breakpoint was close (1–2 bp) to these sites.

The number of seed reads covering the fusion junction is a function of sequencing depth and read length. SRPKM is the number of seed reads normalized by number of mappable reads in the whole NGS dataset and also by the adjusted seed region length. It provided a measure close to the RPKM value of the fusion gene or transcript (Fig. 2E), and facilitated the comparison of abundances of fusion products both between samples and between different read lengths.

3.2 Comparison with FusionSeq

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 $\alpha = 25$ 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: $(75-25*2)*2/(400-75*2) \approx 0.2$ in dataset P1 and $(75-25*2)*2/(158-75*2) \approx 6$ 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 eight threads in total) 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.

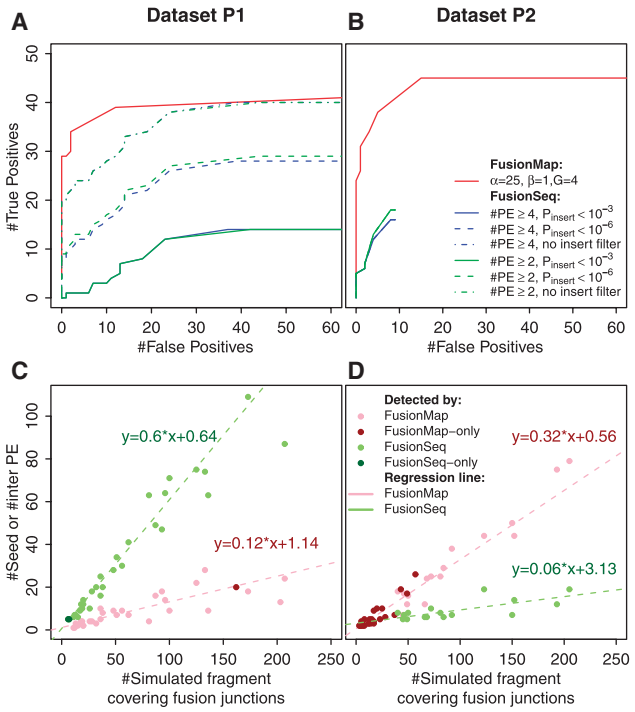

Fig. 3. Comparison of FuionMap with FusionSeq on simulated datasets. (**A** and **C**): dataset P1; (**B** and **D**): dataset P2. (A and B): number of false and true positives detected by two methods. The lines representing the parameter changes for abnormal insert size filter in FusionSeq are overlapped in dataset P2. (C and D): the scatter plot of the number of simulated versus detected fusion reads by both methods. *Y*-axis represents the number of seed reads for FusionMap or the number of inter-transcript read pairs for FusionSeq. The regression lines were computed based on least trimmed squares regression (LTS) on 90% of data points for FusionSeq and FusionMap separately.

Table 3. Computational complexity of fusion detection

Dataset	Method	Performance	
		Time	Peak memory
P1	FusionMap (v1.41)	55 min	6 GB
	FusionSeq (v0.6.1)	> 15 h	5 GB
P2	FusionMap (v1.41)	45 min	6 GB
	FusionSeq (v0.6.1)	> 8.5 h	4 GB

Based on a 64-bit machine with Intel Xeon E5640 (2.66 GHz, two Quad Cores) and 12 GB RAM.

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

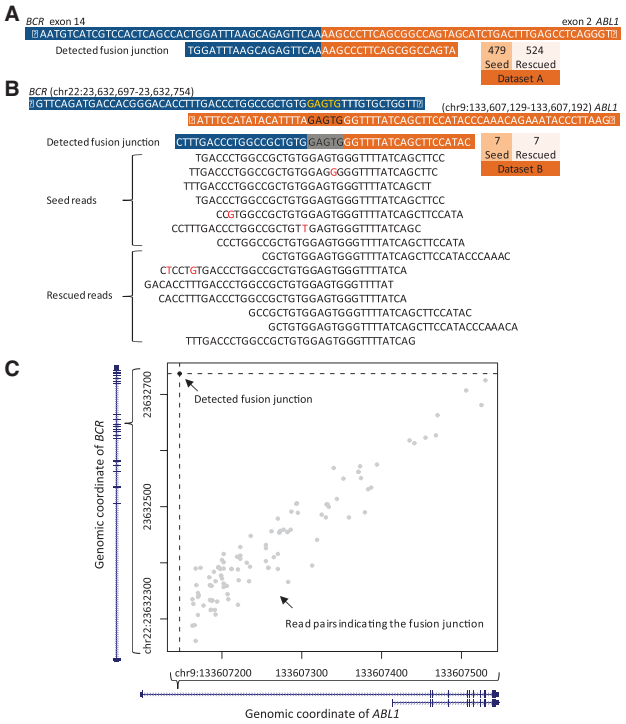

Fig. 4. Detected fusion junctions for *BCR-ABL1* in K562 cell line. (**A**): fusion junctions detected at the transcript level in datasets A. (**B**): detected fusion junctions at the genomic level based on 14 junction-spanning reads in dataset B. (**C**): The scatter plot of chromosomal coordinates of discordant read pairs (grey) and fusion junction-spanning reads (black) linking gene *BCR* and *ABL1*.

major breakpoint cluster region (Quintas-Cardama and Cortes, 2009)] as an anchor and applied Anchored chromosome paired-end tags (ChromPET) technique to capture the genomic sequence of *BCR-ABL1* fusion gene in the K562 cell line (Shibata *et al*., 2010).

For dataset A, we applied FusionMap using $\{\alpha = 25, \beta = 1, G = 2\}$ directly on the whole dataset since it is a single-end dataset. For dataset B, we aligned reads to the human genome hg19 using Bowtie (Langmead *et al*., 2009) allowing two mismatches. A paired-end fusion report was generated based on read pairs with two ends uniquely aligned to different genes. We next applied FusionMap using $\{\alpha = 17, \beta = 1, G = 0\}$ to the unmapped reads to detect fusion junctions. We set $\alpha = 17$ which allowed five possible cutting positions for a 38 nt read.

FusionMap detected *BCR-ABL1* as the top fusion junction at the transcript level in dataset A and it was the sole genomic fusion candidate with more than two supportive seed reads in dataset B. (Fig. 4A and B; Supplementary Table S1). In dataset A, we identified 1003 junction-spanning reads connecting the end of exon 14 in *BCR* to the beginning of exon 2 in *ABL1* at the transcript level. In dataset B, which was from a gDNA-Seq study, we identified 114 pairs of reads linking chr22:23,632,261-23,632,725 in *BCR* and chr9:133,607,163-133,607,530 in *ABL1*, according to the middle positions of mapped reads on human genome hg19 (Fig. 4C). It indicated that the genomic breakpoint in *BCR* is on the 3′ side near chr22:23,632,725 and the breakpoint in *ABL1* is on the 5′ side near chr9:133,607,163. Before the aggregation step, FusionMap detected six fuzzy junctions supported by 14 junction-spanning reads, which showed that the breakpoint of *BCR* was in chr22:23,632,737-23,632,742 and the breakpoint of *ABL1* was in chr9:133,607,147-133,607,152 (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 *BCR* and in the intron region between exon 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 3) 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/(400-75*2) = 0.6. The ratio will increase when the usable read lengths increase with new sequencing techniques or reagents (Illumina, 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 length 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 detecting recurrent fusion events by grouping fusion junctions by rows in one table.

The main limitation of FusionMap is its reliance on long read lengths. 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$75 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

Au,K.F. *et al*. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.

Berger,M.F. *et al*. (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.*, **20**, 413–427.

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Campbell,P.J. *et al*. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Edwards,P.A. (2009) Fusion genes and chromosome translocations in the common epithelial cancers. *J. Pathol.*, **220**, 244–254.

Hampton,O.A. *et al*. (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, **19**, 167–177.

Illumina (2010) SBS sequencing Kit v5 reagent preparation guide.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Levin,J.Z. *et al*. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.*, **10**, R115.

Lugo, T.G., *et al*. (1990) Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science*, **247**, 1079–1082.

Maher,C.A. *et al*. (2009a) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

Maher,C.A. *et al*. (2009b) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Quintas-Cardama,A. and Cortes,J. (2009) Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood*, **113**, 1619–1630.

Rabbitts,T.H. (2009) Commonality but diversity in cancer gene fusions. *Cell*, **137**, 391–395.

Raney,B.J. *et al*. (2010) Encode whole-genome data in the ucsc genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sboner,A. *et al*. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data. *Genome Biol.*, **11**, R104.

Shibata,Y. *et al*. (2010) Detection of DNA fusion junctions for BCR-ABL translocations by Anchored ChromPET. *Genome Med.*, **2**, 70.

Smedley,D. *et al*. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Tkachuk,D.C. *et al*. (1990) Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. *Science*, **250**, 559–562.

Tomlins,S.A. *et al*. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Wang,K. *et al*. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.