Sequence Analysis

FusionScan: accurate prediction of gene fusion from RNA-Seq data

Pora Kim and Sanghyuk Lee*

Department of Life Science and Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 120-750, Republic of Korea

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Identification of fusion gene is of prominent importance in cancer research field because of their potential as carcinogenic drivers. RNA-Seq data have been the most useful source for identification of fusion transcripts. Although a number of algorithms have been developed thus far, most programs produce numerous false positives, which is unacceptable to experimental confirmation. We still lack a reliable program that achieves high precision with reasonable recall rate.

Results: Here, we present FusionScan, a highly optimized tool for predicting fusion transcripts from RNA-Seq data. We specifically search for split reads composed of intact exons at the fusion boundaries. Using 269 known fusion cases as the reference, we have implemented various mapping and filtering strategies to remove false positives without discarding genuine cases. In the performance test using 3 cell line datasets with validated fusion cases (NCI-H660, K562, MCF-7), FusionScan outperformed other existing programs by a considerable margin, achieving the precision and recall rates of 60% and 79%, respectively. Simulation test also demonstrated that FusionScan recovered most of true positives without producing an overwhelming number of false positives regardless of sequencing depth and read length. The computation time was comparable to other leading tools. We also provide several curative means to help users investigate the details of fusion candidates easily. Thus, FusionScan would be a reliable, efficient and convenient program for detecting fusion transcripts that meets the need and standard in the clinical and experimental research.

Availability: Freely available at http://fusionscan.ewha.ac.kr/.

Contact: sanghyuk@ewha.ac.kr

1 INTRODUCTION

Fusion genes are important class of biomarkers in cancer studies. Numerous fusion genes have been established as cancer drivers including BCR-ABL1 fusion in chronic myelogenous leukemia (Kantarjian *et al.*, 2002), TMPRSS2-ERG fusion in prostate cancer (Tomlins *et al.*, 2005), EML4-ALK (Soda *et al.*, 2007) and CD74-NRG1 (Fernandez-Cuesta *et al.*, 2014) fusions in non-small cell lung cancer, and FGFR3-TACC3 in glioblastoma (Singh *et al.*, 2012) and bladder cancer (Guo *et al.*, 2013).

A number of algorithms and programs have been already published for fusion detection problem from RNA-Seq data. Basic idea is to identify the split reads and discordant read pairs that map to two distinct genes. Subsequently, the exact fusion point is determined from the *split* reads where single mate reads overlap the fusion junction, with the fusion-encompassing reads used as supporting evidence. Early approaches following this scheme include FusionSeq (Sboner *et al.*, 2010), ChimeraScan (Iyer *et al.*, 2011), deFuse (McPherson *et al.*, 2011), FusionMap (Ge *et al.*, 2011), TopHat-Fusion (Kim and Salzberg, 2011), and FusionHunter (Li *et al.*, 2011), as extensively reviewed by Zhao and coworkers (Wang *et al.*, 2013).

However, their performance varies dramatically in terms of precision, sensitivity (recall), and computational costs according to the mapping methods, filtering strategies, and parameter optimization. According to recent comparison (Carrara *et al.*, 2013) where the performance of these tools was evaluated using synthetic and experimental datasets, no program showed satisfactory performance. Programs with high sensitivity (ChimeraScan and TopHat-Fusion) predicted thousands of false positives. Programs with low sensitivity (FusionMap, FusionHunter, and deFuse) still produced tens to hundreds of false positives, unacceptable number for experimental confirmation, and had very limited overlap in the results.

Recent programs improved the performance by implementing diverse ideas. FusionQ (Liu *et al.*, 2013) used the concept of residual mapping to extend the short reads. Similarly, BreakFusion (Chen *et al.*, 2012) combined the targeted assembly procedure to overcome the limits owing to short read length. EricScript (Benelli *et al.*, 2012) improved the mapping accuracy by building exon junction reference and recalibration using BLAT (Kent, 2002). Nevertheless, no programs achieved the accuracy over 50% of sensitivity and specificity simultaneously for the experimental datasets. Two programs are notable exceptions even though they have not been tested on public datasets. SOAPfuse (Jia *et al.*, 2013) used a library of fusion junction sequences by partial exhaustion algorithm and a series of filters to enhance confidence. Analyzing two bladder cancer cell lines, they confirmed 15 cases out of 16 predictions, whereas deFuse identified 11 fusions of which 10 were confirmed by RT-PCR experiments. SOAPfusion (Wu *et al.*, 2013) implemented a novel masking and aligning procedure to achieve better sensitivity and false discovery rate than deFuse in the simulation test, but it needs further objective evaluation.

*To whom correspondence should be addressed.

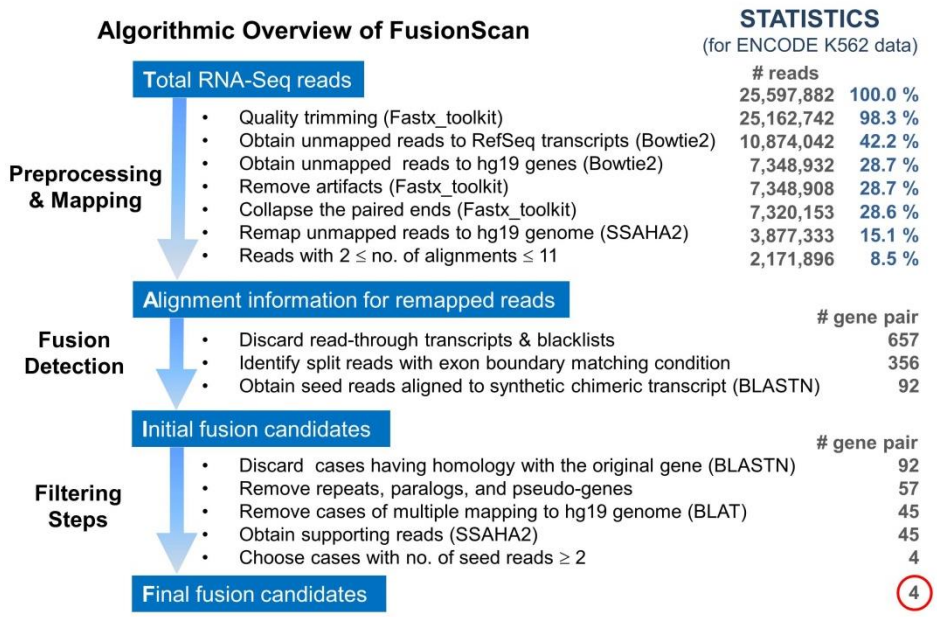

Fig. 1. Overview of FusionScan algorithm and statistics for processing K562 RNA-Seq data

In this article, we report a novel algorithm FusionScan that implemented various strategies to enhance both the sensitivity and precision. We have compared the performance with other widely used programs using both experimental and simulated datasets. Our analysis demonstrated that careful mapping and extensive filtering processes were essential for good performance.

2 METHODS

2.1 FusionScan algorithm

The goal of FusionScan is to identify fusion transcripts composed of combination of intact exons with high sensitivity and specificity. Thus, FusionScan requires multiple *split reads* that join intact exons of two different genes. This may miss cases where the fusion boundary exists inside the exon but the limitation is minor since most of important fusion markers are combination of intact exons thus far. Furthermore, with the advances in sequencing throughput, the read length and sequencing depth of RNA-Seq has became long and deep enough to have multiple split reads including fusion boundaries in most cases.

The algorithm consists of three main parts of preprocessing and mapping, fusion detection, and filtering steps as shown in Fig. 1. Each step is optimized for reliable detection of fusion genes with high sensitivity and specificity as described below. To avoid confusion from naming, we will call two genes involved in the fusion as the *head* and *tail* genes according to the transcription direction of 5'→3', and two exons adjacent to the fusion boundary as *fusion exons*.

2.2 Preprocessing and mapping

Proper preprocessing to identify discordant reads and accurate alignment are the important starting points both for removing reads from normal transcripts for fast processing and for obtaining genuine split reads without loss. We find that these are the critical steps affecting the overall performance that have been overlooked in many cases.

(1) Quality trimming and artifact filtering were done by fastq_quality_trimmer (with the option of '-t 10 –l 38' to keep reads with the minimum length > 38 bp of quality score > 10) and fastq_artifacts_filter in FASTX-Toolkit, respectively (http://hannonlab.cshl.edu/fastx_toolkit/).

(2) Mapping and removing regular reads were carried out in two step procedure. Bowtie2 ver. 2.1.0 (Langmead and Salzberg, 2012) was used to map RNA-Seq reads to the human transcriptome of refGene from the UCSC genome annotation database for the hg19 (GRCh37). Unaligned reads were stored into a file with an option of '-un' and they were realigned to the human genome, further removing reads mapped to the intronic or intergenic regions. Paired-end reads were processed independently in Bowtie mapping to identify discordant split reads. Then, the forward and reverse reads were joined and collapsed using fastx_collapser to produce unique unmapped reads only.

(3) Remapping unmapped reads was achieved by SSAHA2 alignment software (Ning *et al.*, 2001). We have tested several alignment tools for sensitive remapping including GMAP v.2013-11-27 (Wu and Watanabe, 2005), SSAHA2 v.2.5, Bowtie2 v.2.1.0, BWA v.0.7.5a (Li and Durbin, 2009), BLAT v.34, TopHat2 v.2.0.9 (Kim *et al.*, 2013), and MapSplice v.2.1.7 (Wang *et al.*, 2010). For 269 known cases of fusion genes collected from TICdb (Novo *et al.*, 2007) and ChimerDB 2.0 (Kim *et al.*, 2010) (data available in the website), we mapped the synthetic reads of variable length (50 bp, 76 bp, 100 bp) with the fusion break point in the middle of the sequence. SSAHA2 was the best by a narrow margin over Bowtie2 in identifying split reads successfully (Table 1). After extensive testing, we recommend using SSAHA2 with the option

Table 1. Comparison of RNA-Seq alignment programs.

Mapping Program	No. of correct alignments out of 269 known fusion transcripts		
	50bp	75bp	100bp
GMAP	59	28	3
SSAHA2	237	**248**	**252**
Bowtie2	**242**	245	248
BWA	1	238	244
Blat	218	225	226
TopHat2	227	228	226
MapSplice	0	0	242

of '-solexa –skip 6 –cmatch 20 –best 5 –output pslx' to set the seed length as 20 bp and to return five best alignments. Since this is the most time-consuming step, FusionScan supports the multi-thread option to split the unique fasta file and to run SSAHA2 alignment in parallel.

(4) Statistics of preprocessing and mapping are shown in Fig. 1 for the K562 RNA-Seq data of paired-end sequencing from the ENCODE project. Final number of remapped reads that may include the fusion candidates ($2 \leq$ no. of alignments ≤ 11) was reduced to 8.5% of the original data, thus speeding up downstream analysis significantly.

2.3 Fusion detection

FusionScan scans all read alignments looking for split reads whose aligned loci are apart by more than 50 kilo base pairs in the genome. To ensure reliable alignment, we demand the minimum aligned length $\geq$ 20 bp on both sides and that the aligned parts cover more than 50% of the entire read. Two aligned loci of the split read should be contiguous within the range of ±10 bp.

Read-through transcripts are fairly common in the human transcriptome. Co-transcription and intergenic splicing (CoTIS) creates chimeric transcripts connecting exons of two neighboring genes (Communi *et al.*, 2001). Thus, we removed the read-through transcripts between two consecutive genes on the same strand with 5' and 3' ends accordant to the genome annotation. It should be noted that removing read-through cases may remove some genuine gene fusions arising from genomic deletions. We also keep the blacklist of gene fusions, which were removed at this stage. The black list may include unduly frequent gene fusions or fusion predictions that have failed in experimental validation.

Since FusionScan is designed to identify fusion genes specifically composed of intact exons from two participating genes, we apply two steps of examining fusion boundaries. First, we scan the fusion boundaries on the split read so that they appear within 6 bp from the intact boundaries of fusion exons (i.e. exon boundary offset $\leq$ 6 bp) to proceed to the next step. The candidate split read is then realigned to the synthetic sequence of combined fusion exons using the bl2seq tool of BLAST with the word size of 20 (Altschul *et al.*, 1997). Again, the minimum aligned length is 20 bp on both sides with the minimum percent identity $\geq$ 95. Split reads satisfying all conditions given above are the *seed reads* that would strongly support the fusion event. For K562 cell line data shown in Fig. 1, the exon boundary condition and realignment against the synthetic chimeric transcript reduced the number of candidates considerably, and we obtained 92 fusion gene pairs for the filtering procedure.

2.4 Filtering steps

Since most programs for fusion gene prediction yield too many false positives, extensive filtering is essential for reliable performance. In an effort to enhance the precision of the prediction (i.e. small number of false positives), we have implemented several filtering strategies to prevent accidental alignment leading to false split reads as follows:

(1) Homology filter was applied if the nucleotides of 14 bp length before and after the fusion point were homologous to the original sequences of two participating genes. Bl2seq was used to detect homology with the word size of 10.

(2) Filters for repeat regions, paralogs, and pseudo-genes were implemented as well. We discarded the seed reads that were aligned within the repeat regions obtained from the RepeatMasker (Smit *et al.*, 1996) track in the UCSC genome browser. Similarly, gene fusions with paralogous genes obtained from the Duplicated Genes Database (Ouedraogo *et al.*, 2012) or pseudo-genes obtained from the HUGO database (Gray *et al.*, 2013) were removed from the candidates.

(3) In spite of extensive filtering as described above, we still observed many cases where the split read had alternative alignment of similar or better quality elsewhere in the genome. We implemented the multiple mapping filter by running the local version of BLAT v.34 (using the same option as the web version of BLAT) for seed reads to identify such cases of ambiguous multiple mapping (sequence identity > 95%) and removed those from the fusion candidates.

(4) Finally, we choose the fusion candidates with multiple seed reads as reliable (i.e. the minimum number of seed reads = 2).

For K562 cell line data, FusionScan predicted 4 fusion gene pairs in total, and 3 of those were validated experimentally. The workflow in Fig. 1 shows that (i) the homology filter was not effective for this data, (ii) removing repeats, paralogs, and pseudo-genes is an important step of reducing 35 candidates, (iii) recalibration with BLAT alignment is helpful to reduce 12 additional candidates. However, the condition of multiple seed reads was most critical to yield only 4 fusion candidates.

2.5 Curative tools

Even after using various elaborate filters described above, it is often necessary for users to examine the alignment explicitly. We have developed several tools to facilitate visual inspection by users.

(1) Alignment plot is of great help to verify the genuine fusion events. We provide two different types of alignment plot as shown in Fig. 2. Fusion alignment view shows the alignment of fusion reads onto the synthetic fusion sequence. Progressive tiling pattern is the most desirable feature for the genuine fusion genes. Genome alignment view shows the alignment of

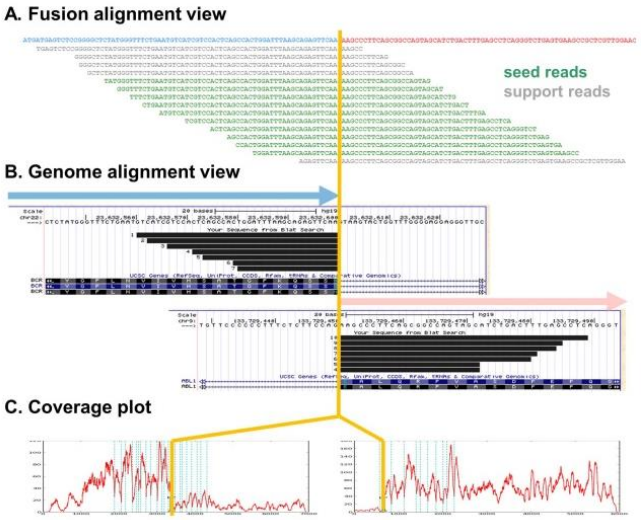

Fig. 2. Alignment and coverage plots for the BCR-ABL1 gene fusion in K562 cell line. Blue vertical lines in C indicate the exon boundaries in the head and tail genes.

fusion transcripts separately for head and tail genes as a custom track in the UCSC genome browser (Fig. 2B).

(2) Coverage plot from NGS data provides valuable information on genomic or transcriptome structures. For example, abrupt depth changes at exon boundaries often indicate the gene fusion or alternative splicing events. FusionScan provides cover age plots for head and tail genes. As shown in Fig. 2C, both the head and tail genes showed abrupt jump at the fusion boundaries in accordance with the fusion event.

(3) Split seed reads are the most direct evidence of gene fusion. In FusionScan, we acknowledge the split reads as the seed only if both sides were aligned to fusion exons over 20 bp long. In cases where only one side met the condition and the other side had a shorter aligned part, we classify them as the *support reads*, which still serve as indirect but good evidence of fusion event. To identify support reads, we realign all RNA-Seq reads to the synthetic chimeric transcripts using SSAHA2 again, and the result is reported with the number of seed reads or used in the fusion alignment plot. This process is optional since it demands realignment of all RNA-Seq reads, taking significant amount of computation. The fusion alignment view may include the support reads as shown in Fig. 2A.

3 RESULTS

Since a number of fusion detection programs are already available in public, it is critical to compare the performance of programs objectively. We have carried out the performance evaluation tests for FusionScan (FS), SOAPfuse v.1.26 (SF), deFuse v.0.6.1 (dF), FusionHunter v.1.4 (FH), FusionMap v.2012-08-12 (FM), and TopHat-Fusion v.2.0.9 (THF) using both experimental and simulation data sets. All programs were run with the default options using the recommended mapping programs and transcriptome model as summarized at the bottom of Table 2. For TopHat-Fusion, we used the output from the TopHat-Fusion-Post that reduced the false positives using BLAST search since it produced too many false positives without the -Post option.

3.1 Comparison of fusion discovery tools using experimental data from 3 cancer cell lines

3.1.1. The data. NCI-H660 is a prostate cancer cell line where two fusion genes (TMPRSS2-ERG and EEF2-SLC25A42) have been verified to play important roles in tumorigenesis. We downloaded the RNA-Seq data from the FusionSeq website (Sboner *et al.*, 2010), which included 6.5 million paired-end reads of 51 bp long.

K562 cell line has long been the standard of leukemia studies where the most famous BCR-ABL1 fusion was identified. Single-end RNA-Seq data for long polyA cytosol mRNAs was downloaded from the Caltech RNA-seq group at the UCSC ENCODE web site. The data includes 12.8 million reads of paired-end sequencing with 76 bp read length. Three cases of gene fusion were known for the K562 cell line (Berger *et al.*, 2010).

One of the most extensively studied samples for gene fusion is the MCF-7 breast cancer cell line. The Caltech RNA-seq group

Table 2. Summary of known fusion genes detected by each tool and the comparison statistics.

Sample	Known (Gold) fusion genes	FS	SF	dF	FH	FM	THF
NCI-H660 (2)	TMPRSS2-ERG	●	●	●	●	●	●
	EEF2-SLC25A42	●	●	●	●		●
	TP/FP	2/0	2/16	2/11	2/1	1/1	2/1
	Precision	1.0	0.11	0.15	0.67	0.50	0.67
	Recall	1.0	1.0	1.0	1.0	0.50	1.0
K562 (3)	BCR-ABL1	●	●	●		●	●
	NUP214-XKR3	●	●	○	●	●	●
	BAT3-SLC44A4	●	●	○		●	●
	TP/FP	3/1	3/7	3/27	1/1	3/12	3/0
	Precision	0.75	0.30	0.10	1.0	0.20	1.0
	Recall	1.0	1.0	1.0	0.33	1.0	1.0
MCF-7 (23)	USP31-CRYL1	●	●	○	○	○	●
	ARFGEF2-SULF2	●	●	○		○	●
	TXLNG-SYAP1	●	●	●		○	●
	DEPDC1B-ELOVL7	●	●	●	○	●	●
	SYTL2-PICALM	●	●	●	○	●	●
	RPS6KB1-DIAPH3		●	●			●
	AHCYL1-RAD51C		●		●	●	●
	TAF4-BRIP1		●			○	
	POP1-MATN2	●	●	●		○	
	GCN1L1-MSI1	●	●	●			●
	ESR1-CCDC170	●	●	●	●	○	●
	SMARCA4-CARM1	●	●	●		○	●
	MYO6-SENP6	●	●	●	●	○	●
	ADAMTS19-SLC27A6	●	●	●	●	○	●
	GATAD2B-NUP210L		●	●		○	●
	SLC25A24-NBPF6	●	●	●		●	
	ATXN7L3-FAM171A2	●	●	●		●	
	C16orf62-IQCK	●	●	●		●	●
	TBL1XR1-RGS17	●					
	BCAS4-BCAS3	●	●	●	●		●
	RPS6KB1-TMEM49	●	●	●		○	●
	ABCA5-PPP4R1L		●				
	C16orf45-ABCC1						
	TP/FP	17/14	21/83	18/132	8/11	17/126	16/37
	Precision	0.55	0.18	0.12	0.42	0.12	0.30
	Recall	0.74	0.91	0.78	0.35	0.74	0.70
Overall	**Precision**	**0.60**	**0.20**	**0.12**	**0.46**	**0.13**	**0.36**
	Recall	**0.79**	**0.93**	**0.82**	**0.39**	**0.75**	**0.75**
	F_1 score	**0.68**	**0.33**	**0.21**	**0.42**	**0.22**	**0.48**
Mapping program		SSAHA2	SOAP2 BWA	GMAP	Bowtie	GSNAP	Bowtie
Transcriptome		RefGene	Ensembl	Ensembl	RefGene	RefGene	Ensembl

- '●' and 'O' indicate that the case was predicted successfully, with direction reversed in 'O'.

- TP = true positive, FP = false positive, Precision = TP/(TP+FP), Recall = TP/(TP+FN),

F_1 score = 2•precision•recall/(precision+recall)

includes RNA-seq data of MCF-7 cell line as well (SRR521521 in SRA database). The data contains 40 million reads of paired-end sequencing with 76 bp read length. Sakarya *et al*. independently studied gene fusions in the MCF-7 cell line using 80 million reads produced by SOLiD paired-end sequencing (Sakarya *et al.*, 2012). They validated 23 gene fusions using TaqMan fusion assays, which were used as gold standards for our benchmark test.

Three data sets of cancer cell lines from public resources represent diverse situations such as different cell types, sequencing depth, single and paired reads, and different read lengths, thus being expected to provide objective result in the comparison test.

3.1.2. Performance comparison. The result from six programs for fusion detection based on RNA-seq data is summarized in Table 2. For fair comparison, we filtered out all cases with the number of seed reads = 1 since FusionScan required the number of seed reads ≥ 2. This may remove some true positives in other programs, but certainly helps in removing false positives. We calculated the precision and recall rates since the true negatives are difficult to prove in gene fusion discovery. It should be noted that we did not penalize other programs for giving wrong direction (i.e. reversed head and tail genes).

In general, the precision and recall rates are contradictory to each other. FusionScan achieved the best in the precision rate (60%) and in the overall performance measured by F_1 score, the harmonic mean of precision and recall rates. SOAPfuse was the best in the recall rate (93%) but its precision rate was just 20%, producing lots of false positives. Fusion-Hunter achieved the precision rate of 46% by sacrificing the recall rate to 39%, missing too many true positives. TopHat-Fusion showed fairly good performance mainly because of recent implementation of extensive filtering scheme in the TopHat-Fusion-Post option.

For experimental biologists or clinicians who carry out validation experiments with limited amount of samples, the precision rate is the most critical attribute. Thus, it is important to note that FusionScan achieved the precision rate of 60% without losing the recall rate considerably (79%). The difference with other programs is substantial, including FusionHunter that achieved excellent performance in recent comparison test by the SOAPfusion study (Wu *et al.*, 2013). It should be noted that one fusion case of C16orf45-ABCC1 was not predicted by all programs, which may suggest that fusion reads for this case were not present in the Caltech RNA-Seq data unlike the SOLiD sequencing data by Sakarya *et al.* Excluding this case, the recall rate of FusionScan increases to 81.5%.

The prediction results from five tools are further illustrated as a Venn diagram in Fig. 3, excluding FusionHunter that missed many true positives. Common hits would have better chance to be genuine fusion cases. FusionScan showed the most common hits from more than three programs (28 out of 31 cases). Importantly, FusionScan had only one singleton prediction, which strongly supports the reliability FusionScan's predictions. FusionMap, deFuse, and SOAPfuse had a number of singleton predictions, most of those being expected to be false positives.

3.2 Comparison of fusion discovery tools using simulation data sets

Testing with experimental datasets is objective and reliable since it reflects diverse situations and experimental conditions that could not be mimicked in simulation studies. However, the scope of benchmark test is limited with small number of known fusion cases and with experimental settings under specific conditions. Thus, we carried out the benchmark test using simulation datasets as well to estimate the performance of each program in different conditions such as variable read length and coverage.

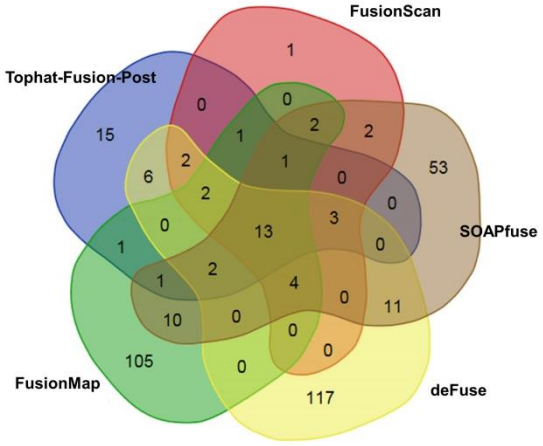

Fig. 3. Venn diagram of fusion predictions for all 3 cell lines. Numbers in parenthesis indicate the total number of predictions.

3.2.1. Preparing the simulation data. Positive cases of fusion gene were artificially constructed by joining two exons of randomly chosen genes, isoforms, and exons in the given order. Adjacent genes were avoided in the selection to exclude read-through transcripts. Using the transcriptome model of refGene, we have generated 10,000 fusion cases for the benchmark test.

For each fusion case, we prepared a synthetic fusion transcript by concatenating the 5' side of the head transcript and 3' side of the tail transcript at the fusion boundary. Random nucleotide position was selected to make a paired end read of desired length (50 bp, 75 bp, or 100 bp) until the pre-determined depth of 10X, 30X, or 50X was achieved. We also demanded the minimum coverage of transcript of 95% (i.e. less than 5% of nucleotides not covered by a sequencing read). The insert size of the paired end reads were selected randomly following the normal distribution with the average insert size of 100 bp and with standard deviation of 10 bp.

Compared to the existing simulation methods that usually add hundreds of synthetic fusion transcripts to the transcriptome model (e.g. RefGene or Ensembl) and run a simulator for producing paired-end sequencing data (Jia *et al.*, 2013; Wu *et al.*, 2013), our procedure of preparing simulation datasets has the advantage of reflecting diverse cases of gene fusion faithfully. The list of 10,000 fusion cases and simulated paired-end sequencing data are available at the website.

3.2.2. Performance comparison. The precision and recall curves from six different programs for fusion detection are shown in Fig. 4 for various sequencing depths and read lengths. Here, we used the default settings of each program for the minimum number of seed reads, instead of demanding two seed reads at least as for the experimental datasets.

TopHat-Fusion-Post showed the highest precision rate consistently but its recall rate was close to 50%. FusionScan was the sec-

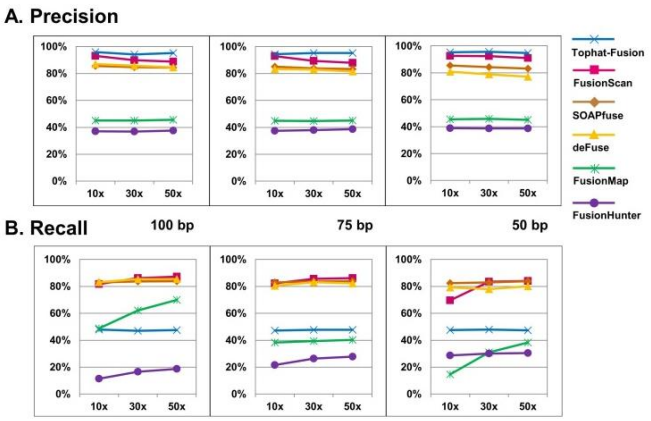

Fig. 4. Precision and recall curves for fusion prediction programs at various read lengths and sequencing depths.

ond to TopHat-Fusion in the precision and the best in the recall rate. At the read length of 100bp and 50X depth, a common practice with recent advances in sequencing technology, FusionScan showed the precision and recall rates of 89% and 87%, respectively. The performance of SOAPfuse and deFuse was slightly inferior to FusionScan in precision and was comparable in the recall rates.

As the sequencing depth increased, the recall rates was improved in all programs. FusionMap and FusionHunter showed substantial variation. The precision rates, however, were fairly independent of sequencing depth and read length.

Overall, the simulated test showed that three programs (FusionScan, SOAPfuse, and deFuse) achieved comparable performance with a slight advantage to FusionScan in the precision. TopHat-Fusion's prediction is reliable, but it misses many true positives as well.

3.3 Implementation and computational resources

FusionScan algorithm was developed using Java (JDK1.7) and Python languages. It further requires many third-party programs such as Bowtie2, SSAHA2, BLAT, bl2seq, samtools and FASTX-Toolkit. Thus, it is highly recommended to run FusionScan in Linux environment with the Java Runtime Environment 1.7 or later.

The CPU time and memory usage are compared in Fig. 5. FusionScan and SOAPfuse took the longest CPU time mainly to achieve the high recall rates. For example, quality trimming with the option of '-t 20 –l 40' instead of '-t 10 –l 38' decreased the run time by half in FusionScan, but lost a few true positives in benchmark testing with 3 cell line datasets. Measuring the CPU time spent for each step of workflow, the preprocessing and mapping took almost half of the total CPU time.

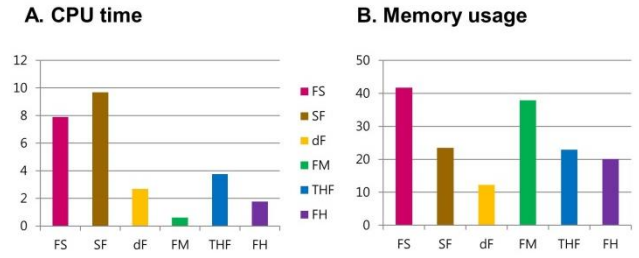

Fig. 5. Comparison of CPU time (hours) and memory usage (GB) for fusion prediction programs.

4 DISCUSSION

For both the real and simulated datasets, the results show that FusionScan provides reliable predictions in fusion discovery under different sequencing coverage and read length.

Even though the general trends were similar between the two datasets, the precision was much worse for the experimental datasets. This indicates that there exist many factors influencing the prediction accuracy in reality. Thus, the result from simulation data should be taken cautiously. Interestingly, SOAPfuse achieved better recall rate for the experimental datasets.

FusionScan was the only program with the precision rate over 50%. The enhanced performance of FusionScan may be ascribed to several points as follows:

(1) Accurate read alignment is absolutely critical. We have selected SSAHA2 as the most sensitive mapping program through a test with known fusion transcripts. This process minimizes the loss of true positives from the start. In a similar effort, SOAPfusion used a special aligner that masked the intronic regions from the transcripts.

(2) Reads with alternative mapping positions should be analyzed cautiously. Many false positives from other programs had their seed reads mapped to other positions concordantly with similar mapping quality. FusionScan removed those ambiguously mapped reads at the filtering steps as described in Fig. 1.

Predicting fusion genes from RNA-Seq data is a procedure full of optimization steps. For example, we have noticed that four true positive cases were filtered out in FusionScan at the final step since they had only one seed read. Relieving the minimum number of seed reads as 1 or using support reads as the basis of rescuing those cases introduced too many false positives.

In conclusion, FusionScan is a reasonable compromise between precision and recall rates, achieving 60% and 79%, respectively, in tests using experimental datasets. With implementation of several curative tools facilitating validation of fusion transcripts, we believe that FusionScan would be a reliable tool for detecting fusion transcripts that meets the need and standard in the clinical and experimental research.

ACKNOWLEDGEMENT

We appreciate Dr. Yeonjoo Jung and Ms. Yeonhwa Jung for carrying out experimental tests on our predictions for MCF-7 cell line.

Funding: The research was supported by grants from the National Research Foundation of Korea (NRF-2014M3C9A3065221, NRF-2012M3A9D1054744, NRF-2012M3A9B9036673), GIST Systems Biology Infrastructure Establishment Grant through ERCSB, and Ewha Global Top5 Grant of Ewha Womans University.

Conflict of Interest: None declared.

REFERENCES

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Benelli, M. *et al.* (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics*, **28**, 3232-3239.

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

Carrara, M. *et al.* (2013) State-of-the-art fusion-finder algorithms sensitivity and specificity. *BioMed Res. Int.*, **2013**, 340620.

Chen, K. *et al.* (2012) BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics*, **28**, 1923-1924.

Communi, D. *et al.* (2001) Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. *J. Biol. Chem.*, **276**, 16561-16566.

Fernandez-Cuesta, L. *et al.* (2014) CD74-NRG1 Fusions in Lung Adenocarcinoma. *Cancer Discovery*, **4**, 415-422.

Ge, H. *et al.* (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, **27**, 1922-1928.

Gray, K.A. *et al.* (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545-552.

Guo, G. *et al.* (2013) Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nature Genet.*, **45**, 1459-1463.

Iyer, M.K. *et al.* (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903-2904.

Jia, W. *et al.* (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.*, **14**, R12.

Kantarjian, H. *et al.* (2002) Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *New Eng. J. Med.*, **346**, 645-652.

Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, **12**, 656-664.

Kim, D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

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

Kim, P. *et al.* (2010) ChimerDB 2.0 - a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81-85.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357-359.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li, Y. *et al.* (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, **27**, 1708-1710.

Liu, C. *et al.* (2013) FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC bioinformatics*, **14**, 193.

McPherson, A. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comp. Biol.*, **7**, e1001138.

Ning, Z. *et al.* (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725-1729.

Novo, F.J. *et al.* (2007) TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, **8**, 33.

Ouedraogo, M. *et al.* (2012) The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PloS One*, **7**, e50653.

Sakarya, O. *et al.* (2012) RNA-Seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comp. Biol.*, **8**, e1002464.

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

Singh, D. *et al.* (2012) Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science*, **337**, 1231-1235.

Smit, A. *et al.* (1996) RepeatMasker Open-3,0. 1996-2000. *<http://www.repeatmasker.org>*.

Soda, M. *et al.* (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561-566.

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

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

Wang, Q. *et al.* (2013) Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief. Bioinformatics*, **14**, 506-519.

Wu, J. *et al.* (2013) SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads. *Bioinformatics*, **29**, 2971-2978.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859-1875.