BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 2971–2978
doi:10.1093/bioinformatics/btt522

Genome analysis

Advance Access publication October 11, 2013

SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads

Jikun Wu[1,†], Wenqian Zhang[2,†], Songbo Huang[1], Zengquan He[2], Yanbing Cheng[2], Jun Wang[2,3,4,5], Tak-Wah Lam[1], Zhiyu Peng[2,6,*] and Siu-Ming Yiu[1,*]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, China, [2]BGI-Shenzhen, Shenzhen, China, [3]Department of Biology, University of Copenhagen, Copenhagen, Denmark, [4]King Abdulaziz University, Jeddah, Saudi Arabia, [5]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen and [6]Guangzhou Key Laboratory of Cancer Trans-Omics Research, BGI-Guangzhou, Guangzhou, China

Associate Editor: Michael Brudno

ABSTRACT

Motivation: RNA-Seq provides a powerful approach to carry out *ab initio* investigation of fusion transcripts representing critical translocation and post-transcriptional events that recode hereditary information. Most of the existing computational fusion detection tools are challenged by the issues of accuracy and how to handle multiple mappings.

Results: We present a novel tool SOAPfusion for fusion discovery with paired-end RNA-Seq reads. SOAPfusion is accurate and efficient for fusion discovery with high sensitivity (≥93%), low false-positive rate (≤1.36%), even the coverage is as low as 10×, highlighting its ability to detect fusions efficiently at low sequencing cost. From real data of Universal Human Reference RNA (UHRR) samples, SOAPfusion detected 7 novel fusion genes, more than other existing tools and all genes have been validated through reverse transcription-polymerase chain reaction followed by Sanger sequencing. SOAPfusion thus proves to be an effective method with precise applicability in search of fusion transcripts, which is advantageous to accelerate pathological and therapeutic cancer studies.

Availability: http://soap.genomics.org.cn/SOAPfusion.html

Contact: smyiu@cs.hku.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 8, 2012; revised on July 13, 2013; accepted on September 2, 2013

1 INTRODUCTION

Fusion transcript, the new transcript transcribed casually from two parental genes, may emerge through chromosomal rearrangements (Kumar-Sinha *et al.,* 2008) or intergenic splicing (Akiva *et al.,* 2006; Horiuchi *et al.,* 2006; Li *et al.,* 2008b). Given their important roles in cancer development and progression (Kantarjian *et al.,* 2002; Kumar-Sinha *et al.,* 2008; Maher *et al.,* 2009a, b; Mitelman *et al.,* 2007; Teixeira *et al.,* 2006), some known fusions have been successfully used as biomarkers for development of inhibitors triggering cancer remission, e.g. *BCR–ABL1* fusion in treating chronic myelogenous leukemia (CML) (Kantarjian *et al.,* 2002). Consequently, it is of enormous biological and therapeutic significance to locate such fusion events.

Early studies mainly resorted to expressed sequence tag (Akiva *et al.,* 2006; Li *et al.,* 2009), array CGH (Shadeo and Lam, 2006) or end sequence profiling (Hampton *et al.,* 2009; Volik *et al.,* 2006) to seek for fusions. However, such methods were largely constrained by their limited sequencing throughput and uneconomic cost (Mortazavi *et al.,* 2008). With the emergence of next-generation sequencing technologies, RNA-Seq has been introduced as an excellent technique in fusion discovery (Mortazavi *et al.,* 2008; Wang *et al.,* 2010), and, in particular, paired-end (PE) sequencing was proved to exhibit distinguished strengths in both productivity and sensitivity (Maher *et al.,* 2009a, b).

The general idea of fusion detection from RNA-Seq data is to align all reads to a reference genome or transcriptome and then explore alignments carrying potential fusion features to call fusions. Early computational detection approaches first narrow down regions containing possible fusions with PE reads intactly mapped to the reference, and then search for the fusion junctions with remaining reads that mapped to those regions intactly. Nevertheless, with successful application of segmental alignment strategy in some splicing junction detection tools, e.g. MapSplice (Wang *et al.,* 2010) and SOAPsplice (Huang *et al.,* 2011), a bunch of computational approaches have been proposed, including FusionSeq (Sboner *et al.,* 2010), ShortFuse (Kinsella *et al.,* 2011), FusionHunter (Li *et al.,* 2011), FusionMap (Ge *et al.,* 2011), deFuse (McPherson *et al.,* 2011) and TopHat-Fusion (Kim and Salzberg, 2011), to consider all segmental alignment of reads in addition to intact alignment.

These current methods vary in performance and ability in fusion detection. Roughly speaking, the performance of these methods rely on their ability to precisely detect the reads encompassing fusion junctions (fusion reads), as those reads served as the major evidences for fusion events and could only be segmentally mapped to references. To obtain specific fusion junctions, current tools align potential fusion reads to combinations of candidate exons (e.g. FusionSeq, ShortFuse, FusionHunter, FusionMap and deFuse) or split those reads into fixed segments for reference alignment (e.g. FusionMap and TopHat-Fusion). This induces higher computational requirement on alignment and/or not easy to solve segmental mapping problems such as

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

multiple mappings. What is more, most of these tools did not perform well under low coverage. As a remark, the existence of multiple mappings (many of them are believed to be incorrect) increases the false discovery rate (FDR; e.g. in ShortFuse, deFuse and TopHat-Fusion). On the other hand, simply removing them decreases the sensitivity as in FusionHunter.

In this article, we present a novel tool called SOAPfusion to identify fusion transcripts with RNA-Seq reads. SOAPfusion integrates a specially designed SOAPfusion-aligner to perform both intact alignment and two-segmental alignment. With a masking strategy on the reference genome and retention of reliable multiple mappings to report fusions, SOAPfusion provides a proper way to make better use of multiple mapping results. Our experiments show that SOAPfusion is an effective tool for detecting fusions at various sequencing coverage even as low as 10×, thus enabling us to identify fusion events at low cost in cancer studies, and in the long run, eventually advance clinical treatment to cancers. In real data, SOAPfusion was able to detect more novel fusion genes than existing tools, in Universal Human Reference RNA (UHRR) samples, all have been validated through reverse transcription-polymerase chain reaction (RT-PCR) followed by Sanger sequencing.

2 METHODS

2.1 System design for SOAPfusion

As the majority of all validated fusions have the fusion sites at exon boundaries (Berger *et al.,* 2010*;* Edgren *et al.,* 2011*;* Levin *et al*., 2009; Sboner *et al.,* 2010), also demonstrated by canonical fusion mechanisms (Akiva *et al., 2006*; Horiuchi *et al.,* 2006; Kumar-Sinha *et al.,* 2008; Li *et al.,* 2008b), the main workflow of SOAPfusion tries to identify fusion transcripts with fusion sites at exon boundaries using RNA-Seq PE reads. In the first step, SOAPfusion masks non-exonic regions of the reference genome with a series of 'N' sequences. Then SOAPfusion uses SOAPfusion-aligner to map all reads to the masked reference genome. In the first alignment stage, all reads are mapped with intact alignment module. Those unmapped ones are continued to the second alignment stage, the two-segmental alignment module. Roughly speaking, there are two types of fusion supporting reads that are required to recover in all fusion detection approaches, the fusion reads (i.e. one of paired reads encompassing fusion site, whereas the other one from either side of fusion site) and the fusion spanning reads (i.e. one read on each side of the fusion site) (Fig. 1A). Thereby, SOAPfusion hunts fusion reads and fusion spanning reads from SOAPfusion-aligner output. Next, SOAPfusion makes use of fusion reads to locate fusion candidates and then calls reliable fusions with supports from fusion spanning reads. Candidates that come from two genes of the same gene family and with none of the fusion supporting reads having unique mappings were filtered out. After that, SOAPfusion reports sets of most reliable fusion candidates. The main workflow is depicted in Figure 1B.

For fusion transcripts with fusion sites from intergenic regions or introns, we use TopHat-Fusion (Kim and Salzberg, 2011) to detect them after the main workflow. In this supplementary step, we use un-masked reference genome and all un-mapped reads (neither intactly nor two-segmentally mapped) returned by SOAPfusion-aligner as input to TopHat-Fusion. In the reported set of fusion candidates by TopHat-Fusion, we only retain those with at least one fusion site in intergenic regions or introns. Fusion candidates from main workflow and supplementary step are included in the final set of reported fusions.

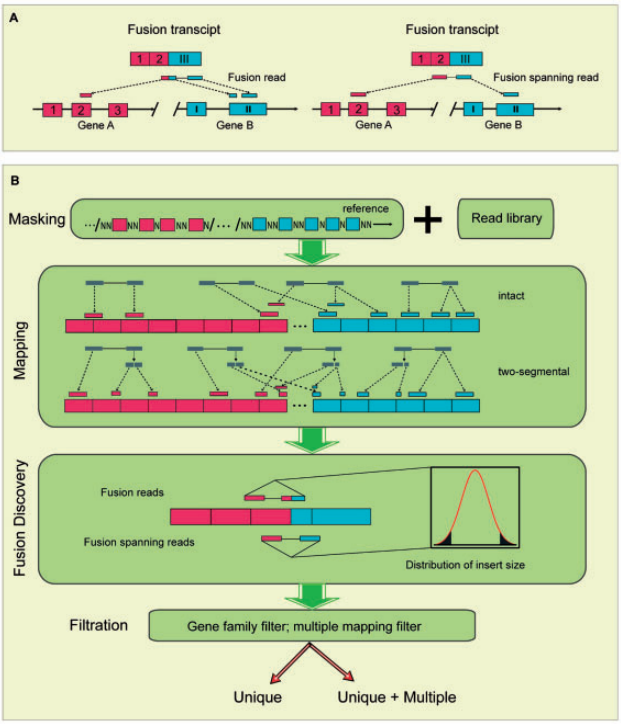

Fig. 1. (**A**) Two types of fusion supporting reads. (**B**) Workflow of SOAPfusion. SOAPfusion first masks non-exon regions on genome reference, then maps PE reads to the masked reference genome with intact alignment and, if necessary, two-segmental alignment, calls fusions with fusion reads and adds fusion spanning reads to enhance confidence. Insert size of fusion supporting reads should follow the normal distribution according to the three-sigma rule (Smirnov and Dunin-Barkovskii, 1969). In the last step, SOAPfusion sifts most reliable fusions after using similarity filters

2.2 Masking algorithm

Based on gene annotation file obtained from UCSC database (http://genome.ucsc.edu), our program keeps all exon regions of human reference genome hg 18 (Build 37) and replaces all nucleotide bases within non-exon regions with 'N's by hard-masking (Baxevanis *et al.,* 2001). In case of overlapping exons among variant transcripts of the same gene, the longest exon as the non-masked region is kept. Then index of masked reference genome is built with bi-directional Burrow Wheeler Transformation algorithm (Lam *et al.,* 2008).

2.3 SOAPfusion-aligner algorithm

In intact alignment, all reads and their reversed complements are aligned to the index, with at most three mismatches (with option -m) or two indels (with option -g) allowed in each alignment by default. Considering that single nucleotide polymorphism occurs more frequently than indels (Li *et al.,* 2008c), a higher penalty score is given to indels (=1) than mismatch (=0.5) in mapping. Mappings with smallest penalty scores are returned. For those less reliable bases at the 3′-end of read owing to the limitations of sequencing technology (Hillier *et al.,* 2008), we cut off 3′ tails of unmapped reads after the above steps by 7 bp (default), and then map the remaining part and their reversed complements to the index allowing errors in the same way as described above. Afterward, for the reads that still fail to be mapped, they will be further mapped with the two-segmental alignment procedure.

In two-segmental alignment, we first get two longest aligned segments in both forward and backward directions, with three requirements: each segment be longer than 8 bp; alignment allows no more than one mismatch and no indels; and alignments must have canonical splice sites (GT/AG). If the two longest mapped sequences overlap, we try the dividing site within the overlapped regions from 3′- to 5′-end until a hit is found; if there is no overlap, the two longest mapped segments would be counted as the best segmentation. In this step, priorities for types of mismatches allowed in two segments are (0, 0) > (0, 1)>(1, 0)>(1, 1) ['>' means better].

2.4 Fusion discovery algorithm

To prepare the most informative mapping results for fusion discovery, we pre-process the outputs from SOAP-fusion module as follows: (i) According to the observation that non-fusion-supporting reads occupy a great portion of the mapping results while they actually are not useful in the fusion discovery stage, we remove non-fusion-supporting mappings from the original set of mapping results, thus reducing greatly the time and memory cost. (ii) Multiple mapping results with high confidence (defined later in the text) are chosen to call candidates by default. If one read in intact alignment or two segments in two-segmental alignment have fewer than six (this value is set empirically, intuitively, if an alignment has only a few mapping positions, the alignment result should be reliable) multiple mappings in total, all its mappings are used in the discovery stage; furthermore, if one segment has multiple mappings, only the mappings with the least number of errors (including mismatch and indels) are kept.

Fusion candidates are first called with fusion reads from two-segmental mappings. It is required that the mapping distance between two ends of fusion read should follow the normal distribution of insert size in sequencing according to the three-sigma rule (Smirnov and Dunin-Barkovskii, 1969). Exact fusion site is calculated according to the mapping position, the mapping orientation and the mapping strand. Then we add evidence to fusion candidates with fusion spanning reads, from both intact alignment and two-segmental alignment results. Mapping distance of these fusion spanning reads should also follow the same rule as fusion reads. Finally, fusion candidates with at least one fusion read and at least one fusion spanning reads supported are considered as reliable ones and are reported in two lists: the first list contains those obtained from unique mappings, whereas the second list contains those obtained from using multiple mappings.

One particular concern at this stage is the calculation of insert size. In case of alternative splicing events, we use a maximum principle, i.e. the insert size is calculated based on the union of all transcripts from the same gene (Supplementary Fig. S2).

2.5 Similarity filters

We remove fusion candidates with homologous gene pairs based on the annotation file from TreeFam database (http://www.treefam.org) and those fusions only supported with multiple mappings.

3 RESULTS

3.1 The data

The melanoma and CML datasets were downloaded from NCBI Gene Expression Omnibus under accession number GSE17593. The breast cancer dataset was downloaded from NCBI Sequence Read Archive under accession number SRA: SRP003186. The UHRR dataset can be retrieved from NCBI SRA database with accession number SRA054573.

3.2 Library construction, sequencing and validation

One microgram of total RNA was isolated from UHRR sample (Agilent, Catalog number 740 000), which comprises 10 human cell lines of mammary gland, liver, cervix, testis, brain, melanoma, liposarcoma, histocyte, T lymphoblast and B lymphocyte, according to the manufacturer's instructions, and was subsequently treated with RNase-free DNase I for 15 min at 37°C to remove residual DNA. Libraries were prepared according to the Illumina's protocol. Poly (A) RNA was isolated using the oligo(dT) beads (Dynabeads® mRNA Purification Kit; Invitrogen, Catalog number 610.06). On chemical fragmentation, double-stranded cDNA was synthesized from these RNA samples using random hexamer–primer and reverse transcriptase (Superscript II; Invitrogen; Catalog number 18 064-014). Following the synthesis of second strand, end repair, adenylate of 3′-ends and adaptor ligation, cDNA was further size-selected on agarose gels (~200 bp). After enrichment of cDNA template by PCR, concentrations and sizes of libraries were measured using DNA 1000 kit (Agilent) on an Agilent 2100 Bioanalyzer, and concentrations were also confirmed by qPCR. Then the PE cDNA libraries were sequenced on HiSeq™ 2000 Sequencing System (Illumina) at 90 bp.

We validated fusion candidates by region-specific PCR amplification of cDNA, which was synthesized from 5 μg of total RNA using the reagents from Invitrogen and TAKARA. Following PCR and gel electrophoresis, all PCR-amplified bands were gel-excised and subjected to Sanger sequencing. All the specific primers and candidate sequences around fusion site are listed in Supplementary Table S9.

3.3 Stepwise evaluation of SOAPfusion

To evaluate the effectiveness of our approach in a full scale, we designed simulation experiments to test the workflow in a stepwise manner.

We simulated 300 fusion transcripts based on 31 399 human RefSeq (Pruitt *et al.,* 2007) transcripts. First, by randomly picking up two transcripts and two exon boundaries of them, we joined the two transcripts to form fusion transcripts at the picked exon boundaries. Also, by randomly picking up one non-exonic site and one exon boundary, we joined them to simulate fusion transcripts with fusion sites from intergenic regions or introns. In total, 300 fusions were simulated. Then we ran MAQ (Li *et al.,* 2008a) on the new transcriptome (including all RefSeq transcripts and the newly simulated fusion transcripts) to simulate PE reads. In total, we generated seven sets of simulated reads under sequencing depth 1×, 5×, 10×, 20×, 30×, 40× and 50×, file sizes of which are 242 Mb, 1.19 Gb, 2.40 Gb, 4.80 Gb, 7.20 Gb, 9.60 Gb and 12 Gb, respectively.

We first assessed the masking strategy by running SOAPfusion with masked reference genome and original reference genome (non-masked), respectively, on 50× simulated dataset (64 150 124 reads in total). Results showed that higher sensitivity and lower FDR were achieved with masking (97.66% and 1.67%) compared with non-masking (89.67% and 3.72%). We checked the read alignments and revealed that the masking strategy (94.51%) mapped more reads than non-masking strategy (89.17%), while all reads mapped with non-masking were covered in masking. To investigate the power of masking, we looked

at the reads that were only mapped with masking. It was found that 85.31% of them could not be mapped to the non-masked genome because they either span more than two exons (33.33%) or contain more sequencing errors (51.98%) than allowed (one mismatch) in segmental mapping. Additionally, we compared cases of multiple mappings between masking and non-masking and found that the number of multiple mapped reads dropped from 4 298 624 in non-masking to 1 210 695 in masking.

Among 410 fusion supporting reads of 28 fusions detected only with masking, 391 (95.4%) reads have unique mappings with masking. We also compared time cost and memory consumption between two strategies. For 64 million reads, masking strategy took 14 h and 0.72 Gb memory, whereas non-masking took 21 h and 5.0 Gb memory. All in all, the advantages of using the masking strategy for improving sensitivity, reducing FDR, saving time and memory are obvious.

Then we measured the performance of SOAPfusion-aligner focusing on two-segmental alignment, given that fusion reads are the keys to fusion discovery and all mapped with segmental alignment. On average, SOAPfusion-aligner mapped 78.10% of fusion reads to the correct fusion junctions under all coverage, while for Bowtie used by deFuse (McPherson *et al.*, 2011), the rate is 55.36% (Supplementary Table S1). Accordingly, SOAPfusion-aligner performs two-segmental alignment more accurately and effectively.

We next evaluated the effect of using multiple mappings in SOAPfusion. Although we used the masking strategy to reduce multiple mapping results to the best extent, there were still some reads mapped to multiple locations because of repetitive regions among some exons. We compared sensitivity and FDR when counting fusions only reported from unique mappings (U) and fusions reported from both unique and multiple mappings (U + M) (Supplementary Table S2). In simulated datasets, when the coverage increases, it is clear that the increases in the number of true fusions (based on U + M) are a lot more than the increases in false fusions reported. By aggregating fusions called from both unique mappings and multiple mappings, SOAPfusion achieved high sensitivity while maintaining reasonably low FDR.

Also, we tested the role of similarity filters in improving SOAPfusion's performance (Supplementary Table S3). For simulated dataset under 50×, 82 of 380 fusions reported were removed, including 55 candidates with parental genes coming from the same gene family and 27 candidates only called from multiple mappings, helping to deduce the FDR from 22.37% to 1.67%.

To summarize, the major reasons that SOAPfusion can outperform other existing tools are as follows: (i) With masking, we are able to increase the mapping ratio and reduce the number of multiple mappings of fusion supporting reads; (ii) our RNA-seq aligner is able to align the fusion spanning reads in an accurate and efficient way; and (iii) by properly handling multiple mappings, we achieve a better balance between retaining more fusion supporting reads and having too many false alignments.

3.4 Integral evaluation of SOAPfusion

3.4.1 On simulated data Having identified critical roles of distinct strategies used in SOAPfusion, we next performed

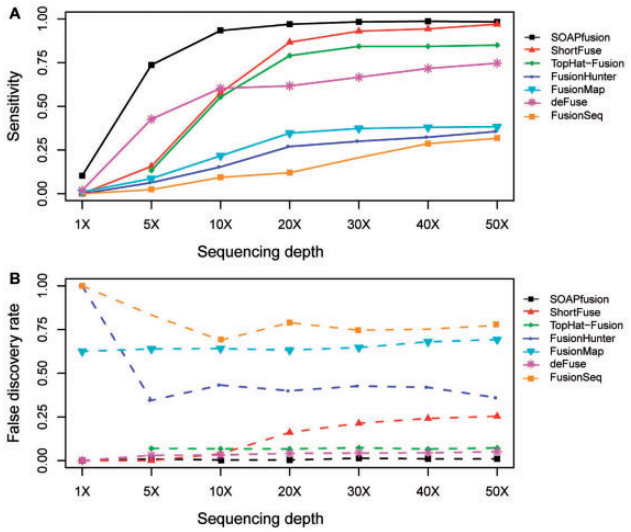

Fig. 2. Performance comparison between SOAPfusion and other tools with simulated datasets. Labels stand for sensitivity (**A**) and FDR (**B**) of each tool with simulated dataset under specific coverage (5× to 50×) confirming the confidence of its predictions

assessments on the whole workflow. As summarized in Figure 2 and Supplementary Table S4, SOAPfusion achieved overall high sensitivity on all datasets except 1×. For datasets under high coverage (20, 30, 40 and 50×), sensitivity of SOAPfusion even approached to 100% (≥97%). One reason for missing some fusions (6 in 20×, 2 in 30×, 2 in 40× and 2 in 50×) is that their fusion reads contained more than three mismatches in segmental mappings, whereas SOAPfusion only allows at most one mismatch for each segment by default. Another reason is that some fusions (3 in 20×, 3 in 30×, 2 in 40×, and 2 in 50×) have fusion reads with more than six multiple mapping positions and thus were skipped in our algorithm. For datasets under extremely low coverage, SOAPfusion had relatively lower sensitivity, obviously in the case of 1× coverage (only 10.33%). From observation, we found that under such low coverage, fewer fusions are covered by fusion supporting reads (Supplementary Fig. S1), and some of them were furthermore missed after mapping if they contained more errors than allowed. Nevertheless, SOAPfusion recovered 73.67% and 93.33% of total simulated fusions under the coverage of 5 and 10×, respectively. Such percentages are acceptable considering the low sequencing cost. Thus, SOAPfusion is efficient in detecting fusions under both high and low coverage. With regard to FDR, SOAPfusion's FDR was <1.36% under all given coverage, demonstrating the accuracy of SOAPfusion.

Moreover, according to Figure 2 and Supplementary Table S4, when sequencing depth increases to >10×, the performance of SOAPfusion is relatively stable in both sensitivity (≥93.33%) and FDR (≤1.36%). We further examined the overlapping set and the unique set of fusions detected among five datasets (10, 20, 30, 40 and 50×), and found that higher depth mainly contributed more fusion supporting reads to fusions in overlapping set, which make them more reliable (Supplementary Table S5), and 30× coverage is already sufficient for discovering 299 of total 300 fusions.

Table 1. Summary of known fusions detected by each tool in melanoma and CML datasets

Sample	5′ Gene	Chr.	3′ Gene	Chr.	SOAPfusion	Short Fuse	Fusion Hunter	DeFuse	Tophat-Fusion	Fusion Map	Fusion Seq
501 Mel	CCT3	1	C1orf61	1	√	√	Y	√	√	Y	–
501 Mel	GNA12	7	SHANK2	11	√	√	Y	–	–	–	–
501 Mel	SLC12A7	5	C11orf67	11	√	√	Y	Y	√	–	–
501 Mel	PARP1	1	MIXL1	1	√	–	√	–	√	–	–
M000216	KCTD2	17	ARHGEF12	11	√	√	√	√	–	√	–
M000921	TMEM8B	9	TLN1	9	√	–	Y	Y	–	–	–
M000921	RECK	9	ALX3	1	√	√	Y	√	√	–	–
M010403	SCAMP2	15	WDR72	15	–	–	Y	–	–	–	–
M980409	GCN1L1	12	PLA2G1B	12	√	–	Y	–	–	–	–
M990802	ANKHD1	5	C5orf32	5	√	√	√	√	√	–	√
M990802	RB1	13	ITM2B	13	√	√	√	Y	√	–	–
K562	BCR	22	ABL1	9	√	√	Y	√	–	√	√
K562	NUP214	9	XKR3	22	√	√	Y	√	–	–	–
K562	BAT3	6	SLC44A4	6	√	–	Y	√	–	–	–

'√' indicates that fusion site, gene pair and their orientations are all correct.
'Y' indicates that gene pair is correct, but the gene orientations are reversed.
'–' indicates that the tool can't detect that fusion.

3.4.2 On published data For further evaluation, we also ran SOAPfusion to rediscover validated fusions from original datasets in previous studies. For melanoma and CML datasets (Berger *et al.,* 2010), SOAPfusion reported 16 fusions (Supplementary Tables S6 and S7), for which the average number of fusion reads and fusion spanning reads were 4.69 and 22.56, respectively; 13 of these fusions have exactly the same sequences as published (Table 1), while one known fusion *SCAMP2-WDR72* failed to be rediscovered. Two PE reads with short segments (8 bp and 11 bp) in two-segmental mapping contained evidence for *SCAMP2-WDR72*, whereas these segments were mapped to >100 positions, and thus discarded by SOAPfusion. For breast cancer datasets (Edgren *et al.,* 2011), SOAPfusion reported 31 fusions (Supplementary Tables S7 and S8), for which the average number of fusion reads and fusion spanning reads were 11.86 and 6.55, respectively; SOAPfusion was able to rediscover 25 of 27 validated fusions (Table 2), while missed three fusions owing to the following reasons: (i) there was no fusion spanning read for candidate fusion *LAMP1-MCF2* and thus it was removed and (ii) there was no fusion supporting read for NFS1-PREX1 fusion and this fusion was not found by all other tools either (Table 2). Note that *ENSG00000236127* is a new gene annotated in hg19 gene annotation, while we used hg18 in this test; after applying hg19 gene annotation, *CSE1L-ENSG00000236127* was correctly recovered. In brief, SOAPfusion is able to detect known fusions.

3.4.3 Novel fusion discovery Although we detected some novel fusion candidates in published datasets, we did not know whether they are correct or not due to the unavailability of biological samples for validation. To determine whether SOAPfusion's ability in detecting novel fusion genes, we sequenced a UHRR sample (Novoradovskaya *et al.,* 2001) and ran SOAPfusion with the reads. In total, SOAPfusion reported seven fusion candidates (Table 3), all of which were validated using RT-PCR followed by Sanger sequencing (Supplementary Table S9). An example is shown in Figure 3, *NPEPPS-TBC1D3* fusion was formed through (17; 17) (q21.32; q12) translocation and its fusion site located at exon 12 (5′-end) of gene NPEPPS and exons 2 (3′-end) of gene TBC1D3. There are a total of 9 fusion reads and 11 fusion spanning reads supporting this fusion.

3.4.4 Comparison of SOAPfusion with other fusion discovery tools To compare SOAPfusion with existing fusion discovery tools, we ran them on both simulated and real datasets.

FusionSeq (v 1.42.1), deFuse (v 0.4.1), FusionHunter (v 1.2), ShortFuse (v 0.2), FusionMap (v 0.6.1) and TopHat-Fusion (v 0.1.0) were used to conduct the comparison. To be consistent, we used reference genome hg18 for all tools. Additionally, to benchmark the performance of different tools, in the first place we set the same threshold for the required number of fusion supporting reads in preliminary experiments. But it was found that this caused some tools to perform far worse than expected according to the statements in their publications. Therefore, we chose to use default parameters or specified parameters suggested in those publications. All tools are tested under CentOS v5.5 on a computer of x86_64 architecture with 30 Gb memory.

Generally, for simulated datasets (Fig. 2), SOAPfusion has the highest sensitivity and the lowest FDR under all coverage; ShortFuse has the second highest sensitivity but high FDR (≥16.13%) under coverage >10×; deFuse has the second highest sensitivity under coverage <10× and the second lowest FDR (≤5.08%) under all coverage; TopHat-Fusion has the third highest sensitivity under coverage >10× and low FDR (≤7.33) under all coverage; FusionHunter and FusionMap have low sensitivities (≤35.67% and ≤38.33%, respectively) and high FDRs (≥34.48% and ≥62.50%, respectively) under all coverage; FusionSeq always performed worse than other tools under all coverage. More details of each tool's performance in simulated datasets are given in Supplementary Table S4.

In real datasets (Tables 1 and 2), SOAPfusion correctly rediscovered the highest number of known fusions (13 of 14 in

Table 2. Summary of known fusions detected by each tool in breast cancer datasets

Sample	5′ gene	Chromosome	3′ gene	Chromosome	SOAP fusion	Short Fuse	Fusion Hunter	deFuse	TopHat-Fusion	Fusion Map	Fusion Seq
BT-474	ACACA	17	STAC2	17	√	√	–	√	√	√	–
BT-474	RPS6KB1	17	SNF8	17	√	√	–	Y	√	–	–
BT-474	VAPB	20	IKZF3	17	√	√	–	Y	√	–	–
BT-474	ZMYND8	20	CEP250	20	√	√	–	Y	√	–	–
BT-474	RAB22A	20	MYO9B	19	√	√	–	Y	–	–	–
BT-474	SKA2	17	MYO19	17	√	√	–	–	√	–	√
BT-474	DIDO1	20	KIAA0406	20	√	–	–	–	–	–	√
BT-474	STARD3	17	DOK5	20	√	–	–	–	–	–	–
BT-474	LAMP1	13	MCF2L	13	–	–	–	Y	–	–	–
BT-474	GLB1	3	CMTM7	3	√	–	Y	–	–	–	–
BT-474	CPNE1	20	PI3	20	√	–	–	–	–	–	–
SK-BR-3	TATDN1	8	GSDMB	17	√	√	Y	Y	√	Y	–
SK-BR-3	CSE1L	20	ENSG00000236127	20	√[a]	–	–	–	–	–	–
SK-BR-3	RARA	17	PKIA	8	√	–	√	√	√	√	–
SK-BR-3	ANKHD1	5	PCDH1	5	√	√	√	√	–	√	√
SK-BR-3	CCDC85C	14	SETD3	14	√	√	–	–	–	–	–
SK-BR-3	SUMF1	3	LRRFIP2	3	√	√	Y	Y	–	–	–
SK-BR-3	WDR67	8	ZNF704	8	√	–	–	–	–	–	–
SK-BR-3	CYTH1	17	EIF3H	8	√	√	Y	Y	–	–	–
SK-BR-3	DHX35	20	ITCH	20	√	–	–	–	–	–	–
SK-BR-3	NFS1	20	PREX1	20	–	–	–	–	–	–	–
KPL-4	BSG	19	NFIX	19	√	√	–	Y	√	–	–
KPL-4	PPP1R12A	12	SEPT10	2	√	√	–	–	–	–	–
KPL-4	NOTCH1	9	NUP214	9	√	√	Y	Y	–	–	–
MCF-7	BCAS4	20	BCAS3	17	√	√	√	–	√	–	√
MCF-7	ARFGEF2	20	SULF2	20	√	–	Y	–	√	Y	–
MCF-7	RP56KB1	17	TMEM49	17	√	√	–	–	√	–	–

'√' indicates that fusion site, gene pair and their orientations are all correct.
'Y' indicates that gene pair is correct, but the gene orientations are reversed.
'–' indicates that the tool cannot detect that fusion.
[a]This fusion was recovered with hg19 gene annotation.

Table 3. Gene fusions reported with SOAPfusion in UHRR (catalog number 740 000)

5′ Gene	3′ Gene	5′ chr	3′ chr	Fusion Read	Support Read	Validation	Literature	Short Fuse	Fusion Hunter	deFuse	Tophat-Fusion	Fusion Map	Fusion Seq
BAT3	SLC44A4	6	6	4	2	√	R1	–	–	–	–	–	–
GAS6	RASA3	13	13	5	3	√	R2	–	–	√	–	√	–
RPS6KB1	TMEM49	17	17	5	2	√	R2	–	–	√	–	Y	–
BCAS4	BCAS3	20	17	6	14	√	R2	√	√	Y	√	Y	–
BCR	ABL1	22	9	4	2	√	R2	√	–	√	√	–	–
ARFGEF2	SULF2	20	20	6	5	√	R2	–	√	Y	–	Y	–
NPEPPS	TBC1D3	17	17	9	11	√	R3	–	–	–	–	–	√

'√' indicates that fusion site, gene pair and their orientations are all correct.
'Y' indicates that gene pair is correct, but the gene orientations are reversed.
'–' indicates that the tool can't detect that fusion.
'R1' refers to (Berger *et al.*, 2010). 'R2' refers to (Maher *et al.*, 2009b). 'R3' refers to (Sboner *et al.*, 2010).

melanoma and CML samples and 24 of 27 in breast cancer samples). FusionHunter can detect all 14 known fusions in melanoma and CML samples, but 10 of them have reversed gene pair orientations; also, it only found 9 fusions in breast cancer samples, with reversed orientations in most of these fusions. ShortFuse, deFuse, TopHat-Fusion and FusionMap all rediscovered fewer known fusions than SOAPfusion with the rediscovery ratios of at most 10 of 14 in melanoma and CML samples and at most 16 of 27 in breast cancer samples. Also, deFuse and FusionMap reported reversed gene pair orientations for some

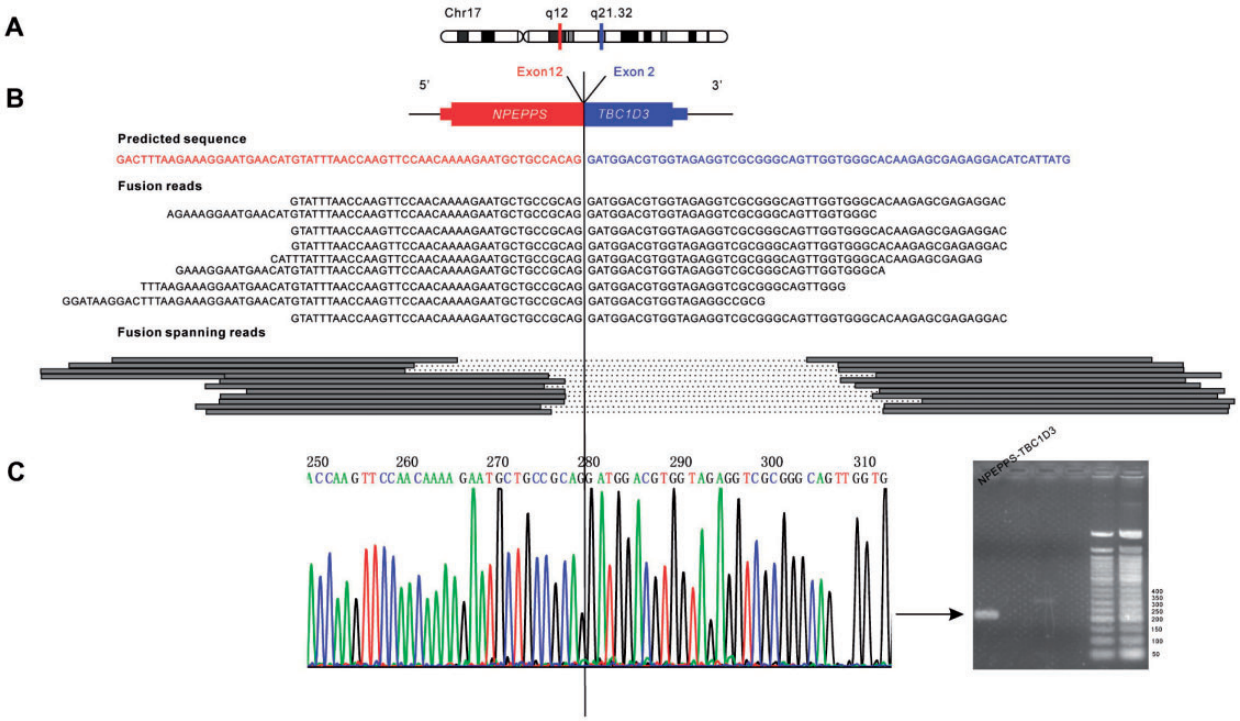

Fig. 3. Illustration of NPEPPS-TBC1D3 fusion. (**A**) genomic locations of two parental genes with 23 exons in NPEPPS and 14 exons in TBC1D3; (**B**) demonstrations of fusion site and the fusion supporting reads including 9 fusion reads and 11 fusion spanning reads. (**C**) RT-PCR and Sanger sequencing validate this fusion

fusions. FusionSeq performed the worst, as it only detected 2 of 14 in melanoma and CML samples and 4 of 27 in breast cancer samples. Moreover, for UHRR datasets, SOAPfusion detected seven true fusions, ShortFuse, FusionHunter and TopHat-Fusion detected only two known fusions; deFuse and FusionMap still predicted fusions with reversed gene pair orientations; FusionSeq detected only one. Details on fusion detection results are listed in Supplementary Table S7.

To conclude, in both real datasets and simulated datasets, the results show that SOAPfusion can provide reliable predictions in fusion discovery under different sequencing coverage.

4 CONCLUSIONS

With adoption of tailor-made aligner for PE RNA-Seq reads, SOAPfusion enables accurate and efficient discovery of fusion transcripts at single-base resolution. SOAPfusion has the highest sensitivity and lowest FDR among existing tools, and it can perform very well even when the coverage is as low as 10×. From real datasets, we also demonstrate its high rediscovery rate and ability in discovering novel fusions. SOAPfusion successfully handles multiple mappings through employment of masking reference genome before alignment and prudent selection of credible multiple alignments in fusion discovery. It was demonstrated that in doing so, on one hand, masking increases the reliability of predictions by retaining much fewer multiple mappings, and retrieves those hard-to-find fusions with fusion reads craggily mapped (>2 segments); on the other hand, with adoption of portion multiple mappings selected in a strict and meticulous way, SOAPfusion has prominent improvement in sensitivity at the expense of just a slight increase in FDR.

SOAPfusion detected and validated a NPEPPS–TBC1D3 fusion, from real dataset UHRR (Fig. 3). We found that both parental genes of NPEPPS–TBC1D3 fusion took part in other fusion events, such as OSBPL9–NPEPPS (Kim *et al.*, 2010), TBC1D3–USP32 (Bailey *et al.*, 2006; Paulding *et al.*, 2003) and NPEPPS–USP32 (Hampton *et al.*, 2009). In fact, NPEPPS gene is an inhibitor of tau-induced neurodegeneration (Karsten *et al.*, 2006). Loss of function mutation on NPEPPS gene will exacerbate neurodegeneration (Sengupta *et al.*, 2006). TBC1D3 gene itself is derived from a segmental duplication (Paulding *et al.*, 2003), with up to eight paralogs resulting in six variant TBC1D3 proteins (Hodzic *et al.*, 2006). *In vitro* studies and analyses of human tumor tissues demonstrated that TBC1D3 expressed widely in tissues (Paulding *et al.*, 2003) and is an oncogene with similar cancerogenic mechanisms of TRE2, encoded by the chimeric fusion of TBC1D3 and USP32 (Hodzic *et al.*, 2006). Accordingly, we infer that NPEPPS–TBC1D3 fusion is a disruptor in brain, which may be of immense importance, and calls for further therapeutic studies.

Of course, SOAPfusion is not a panacea for gene fusion detection. In real situation, besides those described in this study, the mechanism of fusion can be more complicated and may require both biological and computational investigation. Although SOAPfusion now provides an efficient way to *ab initio* dissect gene fusions, facilitating the knowledge of genomic alternation and targets for clinical treatment, fusion discovery is

not a completely solved problem and there is still much space for further improvement and meaningful questions to explore.

ACKNOWLEDGEMENTS

We would like to thank Daniel Nicorici for pointing out that there are now 40 known fusion genes in the breast cancer datasets (see (Kangaspeska *et al*., 2012)) instead of only 27 we used for our benchmark. We will investigate these additional fusion genes in our future work.

Funding: This work was supported by Guangdong Innovative Research Team Program [2009010016, for Z.Y.P.] and General Research Fund of the Hong Kong Government [HKU 719611E and HKU719709E].

Conflict of Interest: none declared.

REFERENCES

Akiva,P. *et al*. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res*., **16**, 30–36.

Bailey,J.A. and Eichler,E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet*., **7**, 552–564.

Baxevanis,A.D. *et al*. (2001) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. 2nd edn. John Wiley & Sons, New York.

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

Edgren,H. *et al*. (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*., **12**, R6.

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

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.

Hillier,L.W. *et al*. (2008) Whole-genome sequencing and variant discovery in *C. Elegans*. *Nat. Methods*, **5**, 183–188.

Hodzic,D. *et al*. (2006) Tbc1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17Q12. *Genomics*, **88**, 731–736.

Horiuchi,T. *et al*. (2006) Alternative trans-splicing: a novel mode of pre-mRNA processing. *Biol. Cell*, **98**, 135–140.

Huang,S. *et al*. (2011) Soapsplice: genome-wide ab initio detection of splice junctions from RNA-seq data. *Front. Gene*, **2**, 46.

Kangaspeska,S. *et al*. (2012) Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms. *PLOS One*, **7**, e48745.

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

Karsten,S.L. *et al*. (2006) A genomic screen for modifiers of tauopathy identifies puromycin-sensitive aminopeptidase as an inhibitor of tau-induced neurodegeneration. *Neuron*, **51**, 549–560.

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–D85.

Kinsella,M. *et al*. (2011) Sensitive gene fusion detection using ambiguously mapping RNA-seq read pairs. *Bioinformatics*, **27**, 1068.

Kumar-Sinha,C. *et al*. (2008) Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, **8**, 497–511.

Lam,T.W. *et al*. (2008) Compressed indexing and local alignment of DNA. *Bioinformatics*, **24**, 791–797.

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.

Li,H. *et al*. (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., **18**, 1851–1858.

Li,H. *et al*. (2008b) Neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, **321**, 1357–1361.

Li,R. *et al*. (2008c) Soap: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Li,X. *et al*. (2009) Short homologous sequences are strongly associated with the generation of chimeric rnas in eukaryotes. *J. Mol. Evol*., **68**, 56–65.

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

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.

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

Mitelman,F. *et al*. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

Novoradovskaya,N. *et al*. (2001) Using universal human reference RNA in microarray gene expression studies. *Nat. Genet*., **27**, 76.

Paulding,C.A. *et al*. (2003) The Tre2 (Usp6) oncogene is a hominoid-specific gene. *Proc. Natl Acad. Sci. USA*, **100**, 2507–2511.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (Refseq): a curated nonredundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.

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

Sengupta,S. *et al*. (2006) Degradation of tau protein by puromycin-sensitive aminopeptidase *in vitro*. *Biochemistry*, **45**, 15111–15119.

Shadeo,A. and Lam,W.L. (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res*., **8**, R9.

Smirnov,N.V. and Dunin-Barkovskii, IV (1969) *Mathematische Statistik in Der Technik*. Deutscher Verlag der Wissenschaften.

Teixeira,M.R. (2006) Recurrent fusion oncogenes in carcinomas. *Crit. Rev. Oncog*., **12**, 257–271.

Volik,S. *et al*. (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res*., **16**, 394–404.

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