SOAPfuse

Jia, Wenlong; Qiu, Kunlong; He, Minghui; Song, Pengfei; Zhou, Quan; Zhou, Feng; Yu, Yuan; Zhu, Dandan; Nickerson, Michael L.; Wan, Shengqing; Liao, Xiangke; Zhu, Xiaoqian; Peng, Shaoliang; Li, Yingrui; Wang, Jun; Guo, Guangwu

Published in:
Genome Biology (Online Edition)

DOI:
10.1186/gb-2013-14-2-r12

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Jia, W., Qiu, K., He, M., Song, P., Zhou, Q., Zhou, F., ... Guo, G. (2013). SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biology (Online Edition), 14(2), [R12].
https://doi.org/10.1186/gb-2013-14-2-r12

Download date: 21. Nov. 2018
SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data

Wenlong Jia1,2†, Kunlong Qiu1,2†, Minghui He1,2†, Pengfei Song2†, Quan Zhou1,2,3, Feng Zhou2,4, Yuan Yu2, Dandan Zhu2, Michael L Nickerson5, Shengqin Wan1,2, Xiangke Liao6, Xiaqian Zhu6,7, Shaoliang Peng6,7, Yingrui Li1,2, Jun Wang1,2,8,9 and Guangwu Guo1,2*

Abstract
We have developed a new method, SOAPfuse, to identify fusion transcripts from paired-end RNA-Seq data. SOAPfuse applies an improved partial exhaustion algorithm to construct a library of fusion junction sequences, which can be used to efficiently identify fusion events, and employs a series of filters to nominate high-confidence fusion transcripts. Compared with other released tools, SOAPfuse achieves higher detection efficiency and consumed less computing resources. We applied SOAPfuse to RNA-Seq data from two bladder cancer cell lines, and confirmed 15 fusion transcripts, including several novel events common to both cell lines. SOAPfuse is available at http://soap.genomics.org.cn/soapfuse.html.

Background
Gene fusions, arising from the juxtaposition of two distinct regions in chromosomes, play important roles in carcinogenesis and can serve as valuable diagnostic and therapeutic targets in cancer. Aberrant gene fusions have been widely described in malignant hematological disorders and sarcomas [1-3], with the recurrent BCR-ABL fusion gene in chronic myeloid leukemia as the classic example [4]. In contrast, the biological and clinical impact of gene fusions in more common solid tumor types has been less appreciated [2]. However, recent discoveries of recurrent gene fusions, such as TMPRSS2-ERG in a majority of prostate cancers [5,6], EML4-ALK in non-small-cell lung cancer [7] and VTI1A-TCF7L2 in colorectal cancer [8], point to their functionally important role in solid tumors. These fusion events were not detected until recently due to technical and analytic problems encountered in the identification of balanced chromosomal aberrations in complex karyotypic profiles of solid tumors.

Massively parallel RNA sequencing (RNA-Seq) using a next-generation sequencing (NGS) platform provides a revolutionary, new tool for precise measurement of levels of transcript abundance and structure in a large variety of species [9-16]. In addition, RNA-Seq has been proven to be a sensitive and efficient approach to gene fusion discovery in many types of cancers [17-20]. Compared with whole genome sequencing, which is also able to detect gene-fusion-creating rearrangements, RNA-Seq identifies fusion events that generate aberrant transcripts that are more likely to be functional or causal in biological or disease settings.

Recently, several computational methods, including FusionSeq [21], deFuse [22], TopHat-Fusion [23], Fusion-Hunter [24], SnowShoes-FTD [25], chimerascan [26] and FusionMap [27], have been developed to identify fusion transcript candidates by analyzing RNA-Seq data. Although these methods were capable of detecting genuine fusion transcripts, many challenges and limitations remain. For example, to determine the junction sites in a given fusion transcript, FusionSeq selected all exons that were potentially involved in the junction from both of the gene pairs, and then covered the exons with a set of ‘tiles’ that were spaced one nucleotide apart [21]. A fusion junction library was constructed by creating all pairwise junctions between these tiles, and the junctions were identified by mapping the RNA-Seq reads to the junction library. This module makes FusionSeq time consuming, especially for genes with more and larger exons. In addition, Fusion-Hunter identified only fusion transcripts with junction sites at the exon edge (splicing junction), but could not
detect a fusion transcript with junction sites in the middle of an exon [24]. Many homologous genes and repetitive sequences often masquerade as fusion events due to ambiguous alignments of short NGS sequencing reads. The lack of effective filtering mechanisms promoted frequent detection of spurious fusion transcripts. Furthermore, several software consumed large amounts of computational resources (CPU time and memory usage), which was a serious problem when analyzing hundreds of samples in parallel.

To address the limitations above, we present a new algorithm, SOAPfuse, which detects fusion transcripts in cancer from paired-end RNA-Seq data. SOAPfuse combines alignment of RNA-Seq paired-end reads against the human genome reference sequence and annotated genes, with detection of candidate fusion events. It seeks two types of reads supporting a fusion event (Figure 1a): discordant mapping paired-end reads (span-read) that connect the candidate fusion gene pairs; and junction reads (junc-read) that confirm the exact junction sites. SOAPfuse applies an improved partial exhaustion algorithm to efficiently construct a putative junction library and also adopts a series of filters and quality control measures to discriminate likely genuine fusions from sequencing and alignment artifacts (Figure 1b; see Materials and methods). The program reports a high-confidence list of fusion transcripts with the precise locations of junction sites in single nucleotide resolution. Furthermore, SOAPfuse supplies the predicted junction sequences of fusion transcripts, which are helpful for the design of bilateral primers in preparation for RT-PCR validation. Moreover, SOAPfuse creates schematic diagrams that can display the alignment of supporting reads (span-reads and junc-reads) on junction sequences and expression levels of exons from each gene pair. Figures are created in lossless image format (SVG, scalable vector graphics) and, with detailed information on fusion events, will facilitate comprehensive characterization of fusion transcripts at single base resolution and will greatly aid manual selection of the fusion events of interest for further research. SOAPfuse can distinguish specific features of RNA-Seq data, such as insert size and read length, so it still works well even when a single sample includes different types of paired-end RNA-Seq data.

**Results**

**Evaluation of performance and sensitivity of SOAPfuse**

To assess the performance and sensitivity of SOAPfuse, we applied SOAPfuse to paired-end RNA-Seq datasets from two previous studies: dataset A, consisting of six melanoma samples and one chronic myelogenous leukemia sample, in which 15 confirmed fusions were detected [19]; and dataset B from four breast cancer cell lines with 27 validated fusions [20]. According to Sanger sequences, we characterized these fusion transcripts using release 59 of the Ensembl annotation database [28], including gene symbols, chromosome locations and exact genomic coordinates of junction sites (Additional file 1). To compare SOAPfuse with other published software (Additional file 2), we also run deFuse [22], TopHat-Fusion [23], Fusion-Hunter [24], SnowShoes-FTD [25] and chimerascan [26] on both RNA-Seq datasets. FusionSeq [21] and FusionMap [27] were abandoned due to computational limitations (Additional file 3). We examined different parameters for each tool to obtain higher sensitivity with lower consumption of computational resources (Additional file 3). For a given fusion event, the distance between a junction site identified by these tools and the real one as determined by previous reports should be less than 10 bp, or the fusion event was considered as not detected. Figure 2 shows the computing resources (CPU time and memory usage) and sensitivity for SOAPfuse and the other five methods (Additional file 4).

For dataset A, which contains approximately 111 million paired-end reads, SOAPfuse consumed the least CPU time (approximately 5.2 hours) and the second least memory (approximately 7.1 Gigabytes) to complete the data analysis (including the alignment of reads against reference), and was able to detect all 15 fusion events. DeFuse and FusionHunter detected comparable numbers of known fusion events (12 to 13 of the 15 fusions), but took 82.1 and 21.3 CPU hours, respectively, at least four times as much as SOAPfuse (Additional file 5). The computational resource cost of SnowShoes-FTD was comparable with SOAPfuse, but SnowShoes-FTD identified only 8 of 15 events. The remaining two tools, chimerascan and TopHat-Fusion, detected four confirmed fusion events but used significantly more CPU hours or memory usage. For dataset B containing approximately 55 million paired-end reads, SOAPfuse detected 26 of the 27 reported fusion events with 4.1 CPU hours and 6.3 Gigabytes memory. The other five tools were able to identify comparable numbers of reported fusions (15 to 21) and cost at least 6.4 hours CPU time. One fusion event, *NFS1-PREX1*, was missed by all methods, including SOAPfuse (Additional file 3).

The process of data analysis for all six tools included two stages: read alignment then detection of fusion events. For both datasets, SOAPfuse, SnowShoes-FTD, and chimerascan consumed less memory than the other three tools. Chimerascan used less memory than SOAPfuse because it used Bowtie [29], which required less memory than SOAP2 [30] in SOAPfuse, to align reads. The memory usage of the other tools (deFuse, Fusion-Hunter, and TopHat-Fusion) were almost two to three times that of SOAPfuse. They reached maximum memory usage at the fusion detection stage, but not at the read alignment stage, which suggests there may still be
**Figure 1** Framework of SOAPfuse for discovering fusion events. (a) Model of the fusion event from Gene A and Gene B that is supported by span-reads and junc-reads. Gene A and reads mapped to it are in blue, and Gene B is in orange. The junction site is marked by a yellow dot. Two span-reads and two junc-reads are shown. (b) The four parts of the SOAPfuse algorithm: read alignment against the human genome reference and annotated transcript sequences; identifying candidate gene pairs by seeking span-reads; detection of predicted fusions; filtering fusions with several criteria to generate the final high-confidence fusion transcript list. Methods with central roles in the algorithm are indicated in red. Steps marked by an asterisk indicate key filtering steps.

**Figure 2** Performance and sensitivity comparison among six tools based on datasets from previous studies. Dataset A is from a melanoma study, and dataset B is from breast cancer research. (a-c) Average CPU time (a), maximum memory usage (b) and sensitivity of fusion detection (c) of six tools are shown in histograms with detailed values (top).
room for algorithm improvement for fusion detection. SOAPfuse uses several optimized algorithms to reduce memory consumption with low cost to computation speed. For the two datasets, SOAPfuse expended less CPU time and memory than most of the other five tools, and reached the highest detection sensitivity, with almost all reported fusion events rediscovered (41 of 42), showing its superior performance and high sensitivity.

**Estimate of the false negative and false positive rates by simulated datasets**

To estimate the false negative (FN) and false positive (FP) rates of fusion detection by SOAPfuse, we applied SOAPfuse to a simulated RNA-Seq dataset. We used the short-read simulator provided by MAQ [31] to generate paired-end RNA-Seq reads from 150 simulated fusions with nine different expression levels (5- to 200-fold; Additional files 3 and 6). We mixed simulated reads with the RNA-Seq dataset (approximately 19 million paired-end reads) from human embryonic stem cells, which was also used as background data by FusionMap [27]. Chimerascan, FusionHunter and SnowShoes-FTD only detected fusion events with junction sites at the exon boundaries. Their performances could not be evaluated because some simulated fusion events harbored junction sites in the middle of exons. We tested deFuse, TopHat-Fusion and SOAPfuse on simulated paired-end reads. Several strategies were applied to fairly compare the performance of these tools (Additional file 3). In total, 149 (99%) of the 150 fusion events were rediscovered, and 142 (94%) were detected by at least two tools, indicating our simulation was reasonable. To be conservative, the performance comparison was based on the 142 events that were supported by at least two algorithms (Figure 3; Additional file 3). In total, 149 (99%) of the 150 fusion events were rediscovered, and 142 (94%) were detected by at least two tools, indicating our simulation was reasonable. To be conservative, the performance comparison was based on the 142 events that were supported by at least two algorithms (Figure 3; Additional file 7).

As expected, FN rates decreased with increasing expression levels of fusion transcripts (Figure 3a). SOAPfuse and deFuse achieved the lowest FN rates at 5% with fusion transcript expression levels of 30-fold or greater. TopHat-Fusion had higher FN rates, especially at low fusion transcript expression levels (5- to 20-fold). For the FP rate (Figure 3b), only SOAPfuse achieved <5% at different fusion transcript expression levels, while deFuse and TopHat-Fusion had higher FP rates at lower fusion transcript expression levels.

Generally, lower FN rates and lower FP rates are contradictory for detection of fusions; however, SOAPfuse and deFuse are good at reducing FN and FP rates during fusion transcript identification. SOAPfuse missed three simulated fusions, which are detected by both deFuse and TopHat-Fusion (Figure 3c), revealing a weakness in analysis of homologous gene sequences and short fusion transcripts of long genes (Additional file 3). In summary, SOAPfuse showed optimal performance with low FN and FP rates at different expression levels of fusion transcripts.

**Application to bladder cancer cell lines**

We next applied SOAPfuse to two bladder cancer cell lines, 5637 and T24. We performed high-throughput RNA-Seq, using Illumina HiSeq sequencing technology, on mRNA from both cell lines and acquired more than 30 million paired-end reads for each (Table 1; see Materials and methods). SOAPfuse identified a total of 16 fusion transcripts, all of which are intrachromosomal and fused at the exon boundaries. We designed primers for RT-PCR experimental validation of all predicted fusions, and Sanger sequencing of the amplicons confirmed 15 (94%) events, of which 6 were detected in both cell lines (Figure 4a; Table 2; Additional file 8). Detailed analysis showed that several confirmed fusion events (Table 2) might be consequences of chromosomal rearrangements. For example, the HADHB-RBKS fusion transcript (Figure 4b) fuses two genes from...
different DNA strands, indicating a potential inversion (Figure S1a in Additional file 9). Furthermore, some fusions implied possible intrachromosomal translocations (Figure S1b in Additional file 9), such as CIRH1A-TMCO7, PSMD8-SIPA1L3, and TIAM1-ATP5O (Figure 4c-e). Intrachromosomal translocations as a mechanism to create fusions were also found in ovarian carcinoma [32] and glioblastoma [33]. To our knowledge, all the

| Sample ID | Read type | Insert size | Read length | Number of paired-end reads |
|-----------|-----------|-------------|-------------|----------------------------|
| 5637      | Paired    | 200         | 90          | 32,228,742                 |
| T24       | Paired    | 200         | 90          | 36,830,100                 |

Table 1 RNA-Seq data from two bladder cancer cell lines

Figure 4 Confirmed fusions in two bladder cancer cell lines. (a) RT-PCR amplifications of confirmed fusions in two bladder cancer cell lines. Marker (M), positive (β-actin) and negative (ddH2O) controls are also shown. Fusion events in red are detected in both cell lines. For fusions that have multiple RT-PCR products, genuine amplicons of a fusion transcript are boxed in yellow. One fusion, SNAP23-LRRC57, reported by deFuse, is further discussed in the text. (b-e) Fusion events that indicate potential chromosomal rearrangements, including potential inversion (b) and intrachromosomal translocations (c-e), are shown. Blue segments are upstream genes, and downstream genes are in orange. Gene symbols are followed with their DNA strands. Exons around the junction sites are drawn with a double slash indicating exons that are not shown. The start positions of upstream genes and end positions of downstream genes are noted with a colon separating chromosomal location and reference genome coordinate. The span-reads and junc-reads from RNA-Seq are shown over and under the junction sequences, respectively. Sanger sequencing of junction sequences are displayed under the junction sites.
Table 2 Confirmed fusion events from two bladder cancer cell lines

| Sample ID | Fusion genes (5’-3’) | Chromosome (5’-3’) | 5’ position | 3’ position | Fusion reads (span/junc) | Detected in both cell lines | Potential chromosomal rearrangement |
|-----------|----------------------|--------------------|-------------|-------------|------------------------|-----------------------------|-----------------------------------|
| 5637      | BDKRB2-BDKRB1        | 14-14              | 96703518    | 96728989    | 2/3                    | Yes                         | No                                |
| 5637      | CRH1A-TMCO7          | 16-16              | 69184807    | 69117388    | 20/20                  | Yes                         | Yes                               |
| 5637      | CLN6-CALML4          | 15-15              | 68521840    | 68489966    | 4/4                    | Yes                         | No                                |
| 5637      | GATSL1-GTF21         | 7-7                | 74867229    | 74145124    | 3/15                   | No                          | Yes                               |
| 5637      | HADHB-BK8K           | 2-2                | 26502983    | 28070964    | 12/10                  | Yes                         | Yes                               |
| 5637      | POLA2-CDC42EP2       | 11-11              | 65063461    | 65082015    | 3/5                    | No                          | No                                |
| 5637      | PSMD8-SIPA1L3        | 19-19              | 38871639    | 38673159    | 5/5                    | Yes                         | Yes                               |
| 5637      | TIAM1-ATPSO          | 21-21              | 32537279    | 35276325    | 9/21                   | Yes                         | Yes                               |
| T24       | BDKRB2-BDKRB1        | 14-14              | 96703518    | 96728989    | 3/3                    | Yes                         | No                                |
| T24       | CRH1A-TMCO7          | 16-16              | 69184807    | 69117388    | 19/24                  | Yes                         | Yes                               |
| T24       | CLN6-CALML4          | 15-15              | 68521840    | 68489966    | 6/5                    | Yes                         | No                                |
| T24       | CTBS-GNG5            | 1-1                | 85028940    | 84967655    | 3/6                    | No                          | No                                |
| T24       | HADHB-BK8K           | 2-2                | 26502983    | 28070964    | 6/7                    | Yes                         | Yes                               |
| T24       | PSMD8-SIPA1L3        | 19-19              | 38871639    | 38673159    | 6/5                    | Yes                         | Yes                               |
| T24       | TIAM1-ATPSO          | 21-21              | 32537279    | 35276325    | 8/28                   | Yes                         | Yes                               |

All information is based on release 59 of the Ensembl hg19 annotation database. Six fusions were detected in both bladder cell lines, and five events may be derived from potential chromosomal rearrangements on the genome. For Fusion reads (span/junc), numbers of span-reads and junc-reads are separated by a slash.

confirmed fusion events have not been reported by previous studies on bladder cancer, indicating their potential significance for further research.

We also used deFuse to reanalyze this dataset and identified 11 fusions, of which 10 (91%) events were able to be confirmed by RT-PCR experiments. Nine of the ten confirmed events were also detected by SOAPfuse (Table S10 in Additional file 8) and the remaining fusion event (SNAP23-LRRC57; Figure 4a) was missed by SOAPfuse. Sanger sequencing shows that exon 5 of SNAP23 is fused to the antisense sequence of exon 4 of LRRC57. This implies that deFuse has a somewhat different definition of a fusion compared to SOAPfuse. After analyzing the characteristics of the fusion transcripts, we found that several novel fusion transcripts that may be derived from chromosomal rearrangements.

In the simulated dataset, SOAPfuse missed three fusion transcripts. The program had some difficulties detecting fusion transcripts from gene pairs having highly similar sequences, and fusion transcripts involving short transcripts of long genes. However, preliminary solutions have been applied to remedy these shortcomings successfully (Additional file 3), and will be included in future versions of SOAPfuse. By analyzing the characteristics of the fusion events, we found that several novel fusion transcripts detected in the bladder cancer cell line datasets were likely to be derived from chromosomal rearrangements of the DNA. Whole genome sequencing will be helpful for determining whether the fusion transcripts are from genomic DNA variations and if the breakpoints can be detected. We have started to develop a new algorithm to detect chromosomal rearrangements that can generate predicted fusion transcripts from whole genome sequencing data based on the results from SOAPfuse. It will be
complementary to SOAPfuse for performing genome analysis of fusions with tools like CREST [34]. We will continuously refine SOAPfuse and update it on our official website.

Conclusions

Here we present an optimized publicly available methodology for identifying novel fusion transcripts from RNA-Seq data. Our results suggest that SOAPfuse achieves better performance than other published tools and it produces a highly accurate list of fusion events in a time-efficient manner. Furthermore, it provides predicted junction sequences and schematic diagrams of fusion events, which are helpful to analyze detected fusions. Overall, SOAPfuse is a useful method that will enable other research groups to make discoveries from their own RNA-Seq data collections.

Materials and methods

Outline of the general approach

SOAPfuse seeks two types of reads (span-reads and junc-reads; Figure 1a) to identify fusion transcripts. Paired-end reads that map to two different genes (a gene pair) are defined as span-reads, and reads covering the junction sites are called junc-reads. Span-reads are used to identify candidate gene pairs, and junc-reads are used to characterize the exact junction sites at single base resolution. Duplicate span-reads and junc-reads are removed before calculating the number of supporting reads (Figure 6a). SOAPfuse contains nine steps in its pipeline (Additional file 10), and can be divided into four parts (Figure 1b): (i) read alignment (steps S01 to S03); (ii) identifying candidate gene pairs (steps S04 and S05); (iii) detection of predicted fusions (steps S06 and S07); and (iv) filtering fusions (steps S08 and S09). A detailed description of the algorithm is in Additional file 3.

Read alignment

SOAPfuse initially aligns paired-end reads against the human reference genome sequence (hg19) using SOAP2 [30] (SOAP-2.21; step S01 in Additional file 10). We divided the reads into three types according to the read alignment results: PE-S01, SE-S01 and UM-S01, where PE stands for paired-end mapped result, SE for single-end mapped result, and UM for unmapped read. PE-S01 reads indicate the paired-end reads mapping to the genome with the proper insert sizes (<10,000 bp), SE-S01 contains paired-end reads in which only one of the two ends mapped to the reference genome, and paired-end reads indicating a fragment with an abnormal insert size or mapped orientation. All unmapped reads are saved in UM-S01 with a FASTA format. PE-S01 is used to evaluate insert size (Additional file 3). SOAPfuse then aligns UM-S01 reads against annotated transcripts (Ensembl release; step S02 in Additional file 10) and generates SE-S02 and UM-S02. To filter unmapped reads caused by small indels, UM-S02 reads are realigned to annotated transcripts using BWA [35] (BWA-0.5.9; maximum number of gap extensions is 5), and the remaining unmapped reads are called filtered-unmapped (FUM).
Iteratively trimming and realigning reads

The latest protocols for NGS RNA-Seq library preparation can generate paired-end reads with an insert size shorter than the total length of both reads (with the 3’ ends of both reads overlapped). The paired-end reads with overlapped 3’ ends may come from the junction regions containing the junction sites and these paired-end reads are not mapped to the reference if the overlapped regions cover the junction sites. These reads are components of FUM generated in step S02 (Additional file 10) and cannot become span-reads, which will reduce the capability of fusion detection. SOAPfuse estimates whether the number of these paired-end reads with overlapped 3’ ends exceeds the threshold (20% of total reads by default). If yes, or the user enables a trimming operation accessible in the configuration file, SOAPfuse will iteratively trim and realign FUM reads to annotated transcripts (Figure 7; step S03 in Additional file 10). The length of reads after trimming should be at least 30 nucleotides (default parameter in SOAPfuse). The trimmed reads that are able to be mapped to annotated transcripts are stored in SE-S03 (Additional file 3). Two steps were used to finish the trimming and realigning operation: first, FUM reads were progressively trimmed off five bases from the 3’-end and mapped to annotated transcripts again until a match was found; second, using the same strategy, we trimmed the remaining FUM reads from the 5’-end. All mapped paired-end reads from these two steps were merged together (step S04 in Additional file 10).

Identifying candidate gene pairs

From all discordantly aligned reads, SOAPfuse seeks span-reads to support candidate gene pairs (step S05 in Additional file 10). Both the span-reads that mapped
Determined the upstream and downstream genes in the fusion events

After obtaining the candidate gene pairs, the upstream and the downstream genes of the fusion were determined based on the information from span-read alignment against the reference. In the process of paired-end sequencing, the fragments are sequenced from bilateral edges to the middle part: one end starts from the 3’ end of the fragment, while the other end starts from the 3’ end of the complementary base-pairing sequence of the fragment (Figure 8a). This information is used to define the upstream and downstream genes in a fusion transcript.

A span-read (paired-end reads ‘a’ and ‘b’) supports a candidate gene pair (Gene A and Gene B). According to the serial number (‘1’ or ‘2’) and mapped orientation (+’ or ‘-’) of paired-end reads (read ‘a’ and ‘b’), there are 16 combinations, but only 4 are rational. These four combinations support two types of fusions in which the upstream and downstream genes are different (Additional file 11, Table S12). The judgment rule is: the gene aligned by reads in the plus orientation must be the upstream gene. Here, we presume that read ‘a’ maps to Gene A and read ‘b’ maps to Gene B (Figure 8b,c). In Figure 8b, read ‘a’ aligns to Gene A (annotated transcripts) in the plus orientation, so Gene A must be the upstream gene; while in Figure 8c, read ‘b’ aligns to Gene B in the plus orientation, so Gene B must be the upstream gene. According to this rule, SOAPfuse defines the upstream and downstream genes in fusion events.

Obtaining the fused regions

Before we defined the fused regions in which the junction sites may located, we obtained a non-redundant transcript sequence from transcript(s) of each annotated gene (Additional file 3). Two methods were used to define the fused regions in gene pairs. In the first method, SOAPfuse bisects each FUM read, and generates two isometric segments, each called a half-unmapped read (HUM read; step S06 in Additional file 10). HUM reads are aligned against candidate gene pairs with SOAP2. A genuine junction read (junc-read) should have at least one HUM read that does not cover the junction site and could map to one gene of the pair. Based on the mapped HUM read, SOAPfuse extends one HUM read length from the mapped position in non-redundant transcripts to define the fused region wherein the junction site might be located (Figure 9a). For HUM reads with multiple hits, all locations of the hits are taken into account. Original reads of mapped HUM reads are called as useful unmapped reads (UUM read).

SOAPfuse also uses span-reads to detect the fused regions in candidate gene pairs (step S07-a in Additional file 10). Span-reads, the paired-end reads supporting the candidate fusion gene pairs, are derived from the fused transcripts and the junction sites are often located in
regions of the fused transcripts between both ends of span-reads. For upstream and downstream genes, we can extend one region with length equal to insert size (evaluated in step S01) from the mapped position of each 3' end span-read to estimate the fused region covering the junction site (Figure 9b). Every gene pair is always supported by at least two span-reads, corresponding to several fused regions that may have overlaps with each other. We presumed that end 1 of a span-read mapped to position MP1 in Gene A, and end 2 of the span-read mapped to position MP2 in Gene B. The lengths of ends 1 and 2 of the span-reads are RL1 and RL2, respectively. The average of insert sizes (INS) and their standard deviation (SD) are evaluated in step S01. The fused regions were estimated by the following intervals:

The intervals of fused regions for the upstream genes are:

\[ [\text{MP1} + \text{RL1} - \text{FLB}, \text{MP1} + \text{INS} + 3 \times \text{SD} - \text{RL2} + \text{FLB} - 1] \]

And the intervals of fused regions for the downstream genes are:

\[ [\text{MP2} + \text{RL2} - \text{INS} - 3 \times \text{SD} + \text{RL1} - \text{FLB}, \text{MP2} + \text{FLB} - 1] \]

In the above formula, a flanking region with length of FLB was considered because sometimes a few bases from the 3' end of a span-read cover the junction sites in the mismatch-allowed alignment.

SOAPfuse combined the fused regions determined by the above two methods to detect the junction sites using the partial exhaustion algorithm as described below.

**Construction of fusion junction sequence library with partial exhaustion algorithm**

To simplify the explanation of the algorithm, we call the fused regions determined by the above two methods as fused regions 1 and fused regions 2, respectively. Fused region 1, defined by the mapped HUM reads, is a small
region covering the junction sites with length smaller than one NGS read. Fused region 2 is a large region defined by the NGS library insert sizes, which are always much longer than HUM reads. Generally, fused region 1 is more useful than fused region 2 to define the junction sites.

However, not all mapped HUM reads are from genuine junc-reads. Sometimes, one unmapped read from a given gene does not map to the gene as a result of more mismatches than are allowed by SOAP2. Unmapped reads like this are not junc-reads and after the bisection into two HUM reads, one of the HUM reads could be...
are generated using Primer (version 5.0) and all primer sequences and expression level of gene pairs (for example, Additional file 12).

To identify the junction sites of fusion transcripts, we used the short-read RNA-Seq technology. First, paired-end sequencing was performed on the Illumina GAIIx system. The raw reads were quality-filtered using a local filtering pipeline (http://www.illumina.com) according to the manufacturer's instructions. The reads were then aligned to the human genome (NCBI Build 36/hg19) using the software Bowtie (version 0.12.6) with default parameters. The alignment results were then analyzed using the software SAMtools (version 0.1.18) to identify reads that mapped to the fusion region.

The predicted junction sequences for further RT-PCR validation were also created, showing the alignments of supporting reads on junction sites. SOAPfuse classifies fused region 2 into two types: overlapped parts between fused regions 1 and 2 to efficiently define the junction events, we removed the initial candidate fusion. The FN and FP rate of SOAPfuse is lower than 0.01.

To validate the fusion transcripts, we carried out this partial exhaustion algorithm to reduce the number of false positives in the junction sequence. The quality of total RNAs was evaluated using an Agilent 2100 Bioanalyzer. The cDNA libraries were sequenced at a sequencing depth of 5- to 200-fold with 90-bp read length using an Illumina GAIIx system. The reads were quality-filtered using a local filtering pipeline (http://www.illumina.com) according to the manufacturer's instructions. The reads were then aligned to the human genome (NCBI Build 36/hg19) using the software Bowtie (version 0.12.6) with default parameters. The alignment results were then analyzed using the software SAMtools (version 0.1.18) to identify reads that mapped to the fusion region.

The predicted junction sequences for further RT-PCR validation were also created, showing the alignments of supporting reads on junction sites. SOAPfuse classifies fused region 2 into two types: overlapped parts between fused regions 1 and 2 to efficiently define the junction events, we removed the initial candidate fusion. The FN and FP rate of SOAPfuse is lower than 0.01.

To validate the fusion transcripts, we carried out this partial exhaustion algorithm to reduce the number of false positives in the junction sequence. The quality of total RNAs was evaluated using an Agilent 2100 Bioanalyzer. The cDNA libraries were sequenced at a sequencing depth of 5- to 200-fold with 90-bp read length using an Illumina GAIIx system. The reads were quality-filtered using a local filtering pipeline (http://www.illumina.com) according to the manufacturer's instructions. The reads were then aligned to the human genome (NCBI Build 36/hg19) using the software Bowtie (version 0.12.6) with default parameters. The alignment results were then analyzed using the software SAMtools (version 0.1.18) to identify reads that mapped to the fusion region.
sequences can be found in Table S11 in Additional file 8. We carried out the RT-PCR amplifications using TaKaRa Taq™ Hot Start Version and performed reactions in 20 μl volumes with 2 μl of 10× PCR buffer (Mg²⁺ Plus), 2 μl of dNTP mixture (each 2.5 mM), 2 μl of primers (each 10 μM), 0.5 μl of TaKaRa Taq HS (5 U/μl), 20 ng of cDNA and up to 20 μl using ddH₂O. The thermocycler program used was the following: (i) 95°C for 4 minutes, (ii) 95°C for 40 seconds, (iii) 55°C to 62°C for 30 seconds, (iv) 72°C for 45 seconds, (v) steps 2 through 4 repeated 35 times, and
Additional material

Additional file 1: Tables S1 - information on all known fusions from two previous studies. Additional detailed information on the known fusions in two previous studies (melanoma and breast cancer researches). All information of fusions is based on release S9 of the Ensembl hg19 annotation database.

Additional file 2: Software selected for evaluation of performance and sensitivity

Additional file 3: Supplementary notes

Additional file 4: Table S3 - detailed information on performance and fusion detection sensitivity of six tools. CPU time, maximum memory usage and sensitivity of fusion detection for each tool are shown. For the multiple process operations, CPU time has been translated to single process usage.

Additional file 5: Table S4 - detection screen of six tools on two previous study datasets

Additional file 6: Tables S5, S6 and S7. Table S5: detailed information on simulated RNA-Seq reads. Table S6: list of 150 simulated fusion events. Table S7: number of fusion-supporting reads for each fusion event.

Additional file 7: Tables S8 and S9. Table S8: TP and FP rates of SOAPFuse, deFuse and TopHat-Fusion based on simulated datasets. Table S9: detailed information on the simulated fusion events detected by SOAPFuse, deFuse and TopHat-Fusion.

Additional file 8: Tables S10 and S11. Table S10: fusion transcripts detected by SOAPFuse and deFuse in two bladder cancer cell lines. Table S11: primers and Sanger sequences of confirmed fusions in two bladder cancer cell lines.

Additional file 9: Information on all known fusions from two previous studies. Additional detailed information on the known fusions in two previous studies (melanoma and breast cancer researches). All information of fusions is based on release S9 of the Ensembl hg19 annotation database.

Additional file 10: Figure S2 - schematic diagrams of nine steps in the SOAPFuse pipeline. The SOAPFuse algorithm consists of nine steps (from S01 to S09) and details of each step are in the Materials and methods or Additional file 3.

Additional file 11: Table S12 - sixteen combination of span-read. There are sixteen combinations based on serial numbers of reads and their mapped orientations, but only four combinations are rational, supporting two types of fusions in which the upstream and downstream genes are different.

Additional file 12: Figure S3 - schematic diagrams of fusion event RECK-ALX3: (a) Alignment of supporting reads against the predicted junction sequence. The upstream part of the junction sequence is in green, and the downstream part is in red. Span-reads are displayed above the predicted junction sequence with the colored dotted line linking paired-end reads. Junc-reads are shown below the junction sequence. (b,c) Expression analysis of the exons in RECK and ALX3 by RNA-Seq read coverage. Transcripts of RECK and ALX3 are shown below the coordinates. The junction site is shown as a red round dot and a green arrow indicates the transcript orientation in the genome sequence. The region covered by the red line is the region mapped by supporting reads. In this case, we found that the expression levels of RECK and ALX3 exons at all sides of junction sites are significantly different. The exons involved in the fusion transcript are expressed more highly than other ones.

Abbreviations
bp: base pair; CPU: central processing unit; FN: false negative; FP: false positive; FUM: filtered unmapped reads; HUM: half unmapped read bisected from FUM reads; NGS: next-generation sequencing; PE: paired-end mapped result; RL: read length; RT-PCR: reverse transcription polymerase chain reaction; SE: single-end mapped result; SVG: Scalable Vector Graphics; UM: unmapped read.

Authors’ contributions
JW and GG conceived and designed the basic algorithm of SOAPFuse. WJ and KQ implemented and optimized the algorithm. PS and SW performed the validation experiments. WJ, KQ, MH, PS, QZ and FZ carried out the comparison among different tools. YY developed the cell lines: XL, XZ and SP tested and deployed the software on TianHe series supercomputers. JW, YL and GG supervised the project and gave advice. WJ, KQ, MH, PS, DZ, MLN and GG wrote and revised the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the National Basic Research Program of China (973 program 2011CB809200, National High Technology Research and Development Program of China (863 Program: 2012AA02A201). This project was also funded by the Shenzhen municipal government of China and the local government of the Yantian District of Shenzhen. Thanks to Xueda Hu (BGI-shenzhen) for giving advice and supporting this project. Thanks to TianHe research and development team of National University of Defense Technology for testing, optimizing and deploying the software on TianHe series supercomputers. We thank D Kim (TopHat-Fusion group), SL Salsberg (TopHat-Fusion group), and A McPherson (deFuse group) for help on operating software. We also thank the FusionMap group for helpful comments on simulation data.

Author details
1 BGI Tech Solutions Co., Ltd, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. 2 BGI-Senzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. 3 School of Life Science and Technology, University of Electronic Science and Technology of China, No.4, Section 2, North Jianshe Road, Chengdu 610054, China. 4 School of Bioscience and Bioengineering, South China University of Technology, Guangzhou Higher Education Mega Centre, Panyu District, Guangzhou 510006, China. 5 School of Computer Science, National University of Defense Technology, No.47, Yanwachi street, Kaifu District, Changsha, Hunan 410073, China. 6 State Key Laboratory of High Performance Computing, National University of Defense Technology, No.47, Yanwachi street, Kaifu District, Changsha, Hunan 410073, China. 7 The Dong Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-1165 Copenhagen, Denmark. 8 The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-1165 Copenhagen, Denmark.

Received: 16 August 2012 Revised: 24 October 2012 Accepted: 14 February 2013 Published: 14 February 2013

References
1. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat Genet 2004; 36:331-334.
2. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007; 7:233-245.
4.Tkachuk DC, Westbrook CA, Andreoff M, Donlon TA, Cleary ML, Suryanarayana K, Homage M, Redner A, Gray J, Pinkel D. Detection of bc-abl fusion in chronic myelogenous leukemia by in situ hybridization. Science 1990, 250:559-562.

5. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XY, Varambally S, Cao X, Simons JF, Gnarra J, Tulliez M, Laxman B, Lee Jung W, Tan CL, sun, Chinnaiyan AM. Recurrent deletion of TMPRSS2 and ETS transcription factor gene in prostate cancer. Science 2005, 306:648-644.

6. Tomlins SA, Laverman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, Yu J, Wang L, Monte L, Kubo M, Sleeve RL, Varambally S, Mehra R, Chinnaiyan AM.

Distinct classes of chromosomal rearrangements create oncogenic ETS fusions in prostate cancer. Nature 2007, 448:595-599.

7. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watabane H, Kurashina K, Hatateka H, Bandou M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sotohana Y, Sugiyama Y, Mano H. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007, 448:561-566.

8. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Dyer Y, Cubillos K, Sougnez C, Voet D, Sarkaria JG, Jing R, Park M, Verhaak RG, Stratton MR, Rout P, Barretina J, Solt DB, Vakiani E, Shao W, Mishina Y, Warmuth M, Jimenez J, Chi Yang D, Signoretti S, Kaelin W, Spardy N, Hahn WC, Yoshida Y, Oskin G, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTTA1-TCF7L2 fusion. Nat Genet 2011, 43:964-968.

9. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptional complexes by RNA-Seq. Nat Methods 2008, 5:621-628.

10. Wang ET, Sandberg R, Luo S, Kuehnbvta I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform isoform regulation in human tissues. Nat Methods 2008, 5:476-476.

11. Hillier LW, Reinke V, Green P, Hint M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transept of c. elegans. Genome Res 2009, 19:657-666.

12. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: Identification of the transcriptional landscape of the yeast genome defined by RNA-Seq. Science 2008, 320:1344-1349.

13. Flicek P, Amode MR, Barrell D, Brent S, Burge C, Cheng Y, Coren J, Cournac L, Dutheil J, Evers D, Grimmond S, et al. Ensembl Genome Annotation Pipeline 2011. Genome Biol 2011, 12:R6.

14. McLaren W, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren RM, Pearson JR, et al. ANNOVAR: for detection and annotation of genomic variants. Bioinformatics 2010, 26:2061-2062.

15. Xu E, Sun M, Karchin R, Zhang Z, Lin J, Shi W, Venter EC, Brown M, Chiu RY, et al. Deep sequencing of human genomic DNA. Proc Natl Acad Sci USA 2010, 107:16868-16873.

16. Wang Y, Li L, Nussbaum R, Gharani M, Chen X, Zhang X, Allen B, et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.

17. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011, 12:R72.

18. Li Y, Chen J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 2011, 27:1708-1710.

19. Sasaki A, Saktivavan A, Handa M, Kariya S, Kariya T, Kariya N, Kariya J, Kariya Y, Kariya K, Kariya H. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.

20. Jia Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011, 12:R6.

21. Stoner A, Habegger L, Pfueger D, Terry S, Chen DZ, Rozovskiy JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demicheli F, Rubin MA, Gerstein MB. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 2010, 11:R104.

22. McPherson A, Harmozodakari F, Zayed A, Guyalany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Olaa MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 2011, 7:e1001318.

23. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011, 12:R72.

24. Li Y, Chen J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 2011, 27:1708-1710.

25. Sasaki A, Saktivavan A, Handa M, Kariya S, Kariya T, Kariya N, Kariya J, Kariya Y, Kariya K, Kariya H. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.

26. Jia Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011, 12:R6.

27. McPherson A, Harmozodakari F, Zayed A, Guyalany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Olaa MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 2011, 7:e1001318.

28. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011, 12:R72.

29. Li Y, Chen J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 2011, 27:1708-1710.

30. Sasaki A, Saktivavan A, Handa M, Kariya S, Kariya T, Kariya N, Kariya J, Kariya Y, Kariya K, Kariya H. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.

31. Jia Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011, 12:R6.

32. McPherson A, Harmozodakari F, Zayed A, Guyalany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Olaa MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 2011, 7:e1001318.

33. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011, 12:R72.

34. Li Y, Chen J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 2011, 27:1708-1710.

35. Sasaki A, Saktivavan A, Handa M, Kariya S, Kariya T, Kariya N, Kariya J, Kariya Y, Kariya K, Kariya H. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.

36. Jia Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011, 12:R6.

37. McPherson A, Harmozodakari F, Zayed A, Guyalany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Olaa MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 2011, 7:e1001318.

38. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 2011, 12:R72.

39. Li Y, Chen J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 2011, 27:1708-1710.

40. Sasaki A, Saktivavan A, Handa M, Kariya S, Kariya T, Kariya N, Kariya J, Kariya Y, Kariya K, Kariya H. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 2011, 39:e100.