

METHOD

Open Access

JAFFA: High sensitivity transcriptome-focused fusion gene detection

Nadia M Davidson^{1*}, Ian J Majewski^{2,3} and Alicia Oshlack^{1,4*}

Abstract

Genomic instability is a hallmark of cancer and, as such, structural alterations and fusion genes are common events in the cancer landscape. RNA sequencing (RNA-Seq) is a powerful method for profiling cancers, but current methods for identifying fusion genes are optimised for short reads. JAFFA (<https://github.com/Oshlack/JAFFA/wiki>) is a sensitive fusion detection method that outperforms other methods with reads of 100 bp or greater. JAFFA compares a cancer transcriptome to the reference transcriptome, rather than the genome, where the cancer transcriptome is inferred using long reads directly or by *de novo* assembling short reads.

Background

Chromosomal rearrangements have the potential to alter gene function in many different ways; for example, they may produce chimeric fusion proteins that gain new functionality, or place a gene under the control of alternative regulatory elements [1,2]. Fusion genes including BCR-ABL, PML-RAR and EML4-ALK have become targets for therapy in cancer, and as a result there is great interest in defining the full complement of oncogenic fusion genes.

Next generation sequencing of RNA (RNA-Seq) has greatly accelerated the discovery of novel fusion genes in cancer [3-5]. However, while a large number of tools have been presented to identify fusion event using RNA-Seq [5-9], practical use of fusion finding tools is often hampered by either a high false detection rate or low sensitivity [10,11]. Many fusion detection methods identify transcriptional breakpoints by splitting short reads into even shorter segments and then aligning these segments to the genome [5,12]. Short read sequences have lower alignment specificity particularly in the presence of SNPs, sequencing errors and repeat regions. Incorrect mapping of these short read fragments has the potential to lead to false predictions. To overcome this, algorithms look for supporting information, such as neighbouring

reads, or read pairs, that cover the same breakpoint. This strategy can be effective at controlling the false discovery rate, but often requires restrictive filtering that may limit sensitivity.

Another limitation of many fusion finding algorithms is that they have been built and tested using reads shorter than 100 bp. Sequencing reads are becoming longer, with 100 bp paired-end reads now standard for many applications, and read lengths are continuing to increase. The MiSeq and PacBio platforms already produce reads of several hundred and several thousand bases, respectively. It is not clear how current fusion finding algorithms will perform on long read data. For example, many will not work on long single-end data, because they require paired-end reads.

In this study we outline a new method for detecting fusion genes that can be applied to any read length, single or paired-end. A critical and unique feature of our method is that rather than comparing a tumour transcriptome to the reference genome we compare it to the reference transcriptome. There are several advantages in alignment to the transcriptome rather than genome; the complexity of splice site alignment, which can be error prone [13,14], is avoided as the transcriptome only includes exonic sequence; identifying fusion transcripts from those alignments is simplified because we do not need to check if the break can be explained by splicing; and finally, the reference transcriptome consists of less sequence than the reference genome, allowing for slower, but more accurate alignment algorithms to be used, such as BLAT [15]. Critically, BLAT works well

* Correspondence: nadia.davidson@mcri.edu.au; alicia.oshlack@mcri.edu.au

¹Murdoch Childrens Research Institute, Royal Children's Hospital, Flemington Road, Parkville, Victoria 3052, Australia

⁴Department of Genetics, The University of Melbourne, Parkville, Victoria 3010, Australia

Full list of author information is available at the end of the article

over a range of reads lengths, whereas mapping algorithms used by other fusion finders are optimised for short reads. For example, bowtie [16], the recommended aligner for TopHat-Fusion [6], will not map reads longer than 1,024 bases.

Our new method, called JAFFA, is designed for detecting fusions in RNA-seq data with contemporary read lengths. Fusions may be identified using reads from 100 bp up to full-length transcripts. Reads shorter than 100 bp can be analysed effectively by assembling them *de novo* into contigs of 100 bp or longer - a step which is performed by JAFFA. Hence, JAFFA is a complete pipeline; it uses *de novo* assembly or raw reads directly to align to a reference transcriptome and outputs candidate fusions along with associated information such as the position of the break in the genome, a prediction of reading frame, read support metrics and whether the fusion is present in the Mitelman database [17]. JAFFA also reports the sequence of the fusion read or assembled contig. JAFFA is built using the Bpipe platform [18] and takes advantages of features such as modularity of the pipeline stages, running numerous samples in parallel, and integration with computing clusters. JAFFA is therefore a highly effective tool for large RNA-Seq studies involving multiple datasets and samples. The idea behind JAFFA has already been used to successfully identify fusions in lung cancer [19].

We validated JAFFA on a range of data with different read-lengths, including 50 bp, 75 bp, 100 bp and 250 bp paired-end reads as well as ultra-long PacBio reads [20,21]. We used RNA-Seq from breast cancer cell lines [22], glioma tumours [23] and simulation, and found JAFFA has a low false discovery rate without compromising on sensitivity. JAFFA may be run in three defined modes: assembling short reads (shorter than 60 bp), using long reads directly (100 bp or greater), or a hybrid approach that both assembles and processes unmapped reads (between 60 bp and 100 bp). We performed a detailed comparison to established methods and found that JAFFA consistently gave the best performance on contemporary data with reads longer than 50 bp. On 100 bp datasets and longer, JAFFA's computational requirements were comparable to those of other fusion finding tools.

Methods

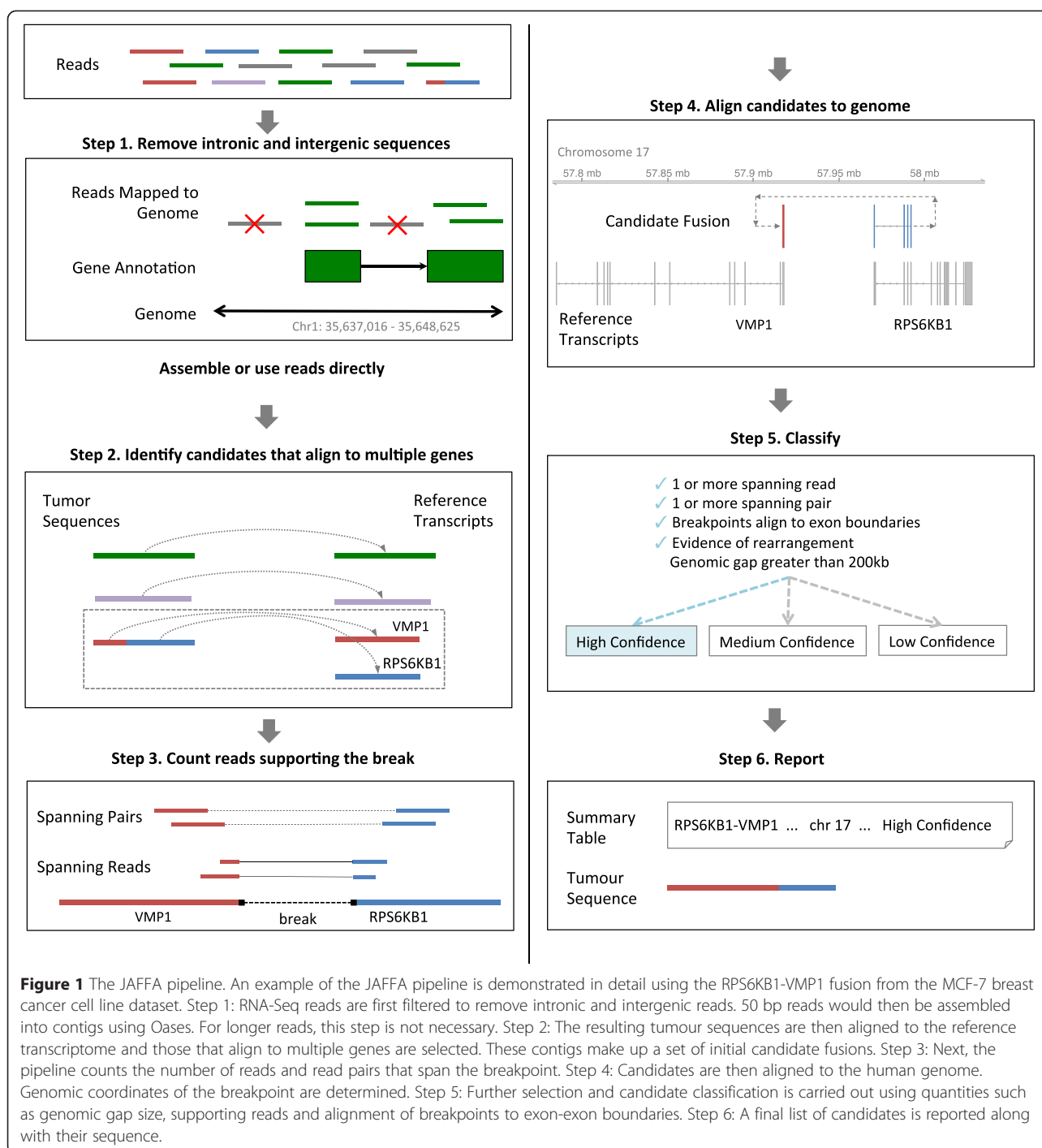
The JAFFA pipeline

JAFFA is a multi-step pipeline that takes raw RNA-Seq reads and outputs a set of candidate fusion genes along with their cDNA breakpoint sequences. JAFFA runs in three modes: (1) 'Assembly' mode assembles short reads into transcripts prior to fusion detection; (2) 'Direct' mode uses RNA-Seq reads directly, rather than assembled contigs, by first selecting reads that do not map to known transcripts; or (3) 'Hybrid' mode both assembles

transcripts and supplements the list of assembled contigs with reads that do not map to either the reference transcriptome or the assembly. The appropriate mode to use depends on the read length (Additional file 1: Figure S1). By default, JAFFA requires 30 bases of flanking sequence either side of the breakpoint. For reads shorter than 60 bp, the flanking sequence would be too short to accurately and efficiently align using BLAT, so the Assembly mode must be used. For reads 60 to 99 bp long, Hybrid mode is used, while for reads 100 bp and over there is no advantage in performing a *de novo* assembly so the Direct mode is used. When *de novo* assembly is performed, Oases [24] is used. We found Oases gave superior sensitivity compared with other assemblers (Additional file 1: Material 1, Additional file 2). *De novo* assembly is well known to producing a high fraction of false chimeras [25,26] and we found an effective method to control for these by checking the amount of sequence shared by fusion partner genes at the breakpoint (Additional file 1: Material 1, Additional file 1: Figure S2).

JAFFA is based on the idea of comparing a sequenced transcriptome against a reference transcriptome. As a default, JAFFA uses transcripts from GENCODE [27] as a reference. For all JAFFA modes, reads aligning to intronic or intergenic regions are first removed to improve computational performance (step 1 in Figure 1). Sequences are then converted into a common form - tumour sequences - consisting of either assembled contigs or the reads themselves. These sequences are processed by a core set of fusion-finding steps (steps 2 to 6 in Figure 1). First, sequences are aligned to a reference transcriptome and those that align to multiple genes are selected. Second, read support is determined. Third, putative candidates are aligned to the genome to check the genomic position of breakpoints. Finally, JAFFA calculates characteristics of each fusion and uses this to prioritise candidates for validation. Each of these pipeline steps is described in detail below.

Most fusion genes originate from a genomic rearrangement with breakpoints in intronic DNA. We found empirically that transcriptional breakpoints aligning to exon-exon boundaries were more indicative of a true fusion than the number of reads supporting the breakpoint, and have incorporated this into our ranking system. Genes with breakpoints aligning to exon-exon boundaries are classified as either 'High Confidence' or 'Medium Confidence'. These two categories are distinguished by either the presence ('High Confidence') or absence ('Medium Confidence') of both spanning reads and spanning pairs. Spanning reads have the fusion breakpoint sequenced within the read. Spanning pairs lie on opposite sides of the breakpoint (Step 3 in Figure 1). For single-end data, only 'Medium Confidence' is reported because spanning pairs are not calculated. Unlike



other fusion finding algorithms, such as deFuse and TopHat-Fusion, which apply a threshold on the number of supporting reads to ensure the false discovery rate is controlled, JAFFA can detect fusions with a single read, without compromising the false discovery rate. Fusions with spanning pairs, but without transcriptional breakpoints aligning to exon boundaries are classified as 'Low Confidence'. For 'Low Confidence' fusions we require

two spanning reads so that chimeric artifacts produced during library preparation are removed. Fusions without spanning pairs or breakpoints aligning to exon boundaries are discarded. Finally, JAFFA flags a fourth class of candidates 'Potential Regular Transcript', which appear to be novel transcripts between adjacent genes [28]. We identify these by a genomic gap between the breakpoints of less than 200 kb and no evidence for genomic

rearrangement. Because these candidates are likely to be caused by read-through transcription [29], they are excluded from the default reporting of our software. For candidates within a class, we rank by the sum of spanning reads and spanning pairs. When read support is equal, we rank on the genomic gap size, with smaller gaps ranked higher as we found empirically that true positives were often intrachromosomal and localised (Additional file 1: Figure S3).

Because JAFFA is a pipeline rather than a standalone software tool, many of its stages rely on external software. The choice of these programs, the reference annotation and genome can be easily customised. In JAFFA, bash and R scripts are used to steer each step, and the pipeline is implemented using the Bpipe platform [18]. Bpipe handles parallelisation, restarting from midway through the pipeline and error reporting, and is convenient for analyses involving a large number of samples. Below, we describe each stage of JAFFA version 1.06 in more detail along with the software choices used during validation. JAFFA is open source and available for download from [30].

Preliminary read filtering

To aid in computation efficiency, JAFFA begins by filtering out any reads that map to intronic, intergenic or mitochondrial sequence in the genome. This is achieved through a two-step process. Initially all read pairs that map concordantly to the reference transcriptome will be retained. Those that do not map, will move to the second step, where they will be mapped to a version of the human genome, hg19, with exonic sequence masked out. Any read pairs that fail to map concordantly will be retained and merged with those from the initial step. Approximately, 70% to 95% of reads pass this filter.

Assemble reads

Short reads were *de novo* assembled using Velvet version 1.2.10 and Oases version 0.2.08 with k-mer lengths of 19, 23, 27, 31 and 35. We required Oases to output contigs with 100 bases or more. Other settings were default.

Remove duplicates

BBMap version 33.41 [31] was used to remove duplicate reads and convert the fastq reads to fasta format.

Select reads that do not map to known transcripts

In the case of the Direct mode, reads were mapped as single-end to sequences from GENCODE version 19. We used bowtie2 with the option '-k1 -un' for the alignment. For the Hybrid mode, we mapped reads to the GENCODE transcriptome, then took the reads that did not map and attempted to map these to the *de novo*

assembled transcriptome. The same bowtie2 settings as above were used.

Align contigs/reads to known transcripts

We used BLAT [15] to align transcript sequences. When aligning to the transcriptome, we required 98% sequence identity over more than 30 bases, with no intronic gaps, '-minIdentity = 98 -minScore = 30 -maxIntron = 0'. A tile size of 18 was used to improve computational speed, '-tileSize = 18', for the assembly mode, or for reads longer than 100 bp, otherwise a tile size of 15 was used to improve sensitivity. These BLAT options are the default in the JAFFA pipeline.

Select contigs/reads that match multiple genes

We first did a loose selection step to identify which tumour sequences aligned to multiple reference transcripts. The two (or more) reference transcripts were required to be separated by 1 kb in the genome by default. Following this we calculated the number of bases that the reference transcripts had in common at the breakpoint. If two genes contained the same sequence over a length that was more than the minimum assembly k-mer length (19 bases), a false chimera may be reported. We controlled for this by only selecting fusion candidates with 13 bases or less of sequence in common between the reference genes (Additional file 1: Figure S2). This step was implemented as an R script.

Counting reads and pairs spanning breakpoints

We counted the number of spanning reads and spanning pairs across the breakpoint. Spanning reads were defined as reads that lay across the breakpoint. Spanning pairs were defined as pairs in which the reads of each pair, lay in their entirety, on opposite sides of the breakpoint. This calculation was performed differently depending on whether the reads were assembled or not. For assembled reads, the reads were mapped back to the candidate *de novo* transcript sequences using bowtie2 with the alignment flags of '-k1 -no-unal -no-mixed -no-discordant'. Spanning reads were required to have 15 base pairs of flanking sequence either side of the break. For the direct mode, spanning pairs were calculated by mapping reads to the reference transcriptome and searching for discordantly aligned pairs, consistent with the predicted fusion. Each fusion candidate in Direct mode was initially assigned one spanning read (that is, since the sequence for which the candidate was identified was itself a read). Therefore in this mode, the minimum flanking sequence was 30 bp, the minimum to identify a fusion. When multiple reads or contigs predicted the same breakpoint the read support was aggregated. Note that spanning pairs will not map when the break lies close to the beginning or end of a transcript.

Aligning candidate contigs/reads to the genome

We aligned the candidate fusion sequences to the human reference genome (hg19) using BLAT with default options.

Check genomic gap, frame and classify candidates

The genomic coordinates of each breakpoint were found and the genomic gap size calculated. In some cases, the gap was very small (less than 10 kb) indicating that the candidate was likely to be a false positive, generally due to families of genes with similar sequence or repeated sequence in the genome. These candidates were discarded. Candidates between adjacent genes can also be reported due to run-through transcription or unannotated splicing. We tried to distinguish these scenarios from genuine fusions with small gaps, by looking for evidence of a genomic rearrangement or inversion, based on the direction of the *de novo* transcript with respect to the genome. If no such evidence was found and the gap was less than 200 kb the fusion was flagged as a 'PotentialRegularTranscript' (not reported by default). Next we determined whether the breakpoints lay on known exon-exon boundaries, as would be expected if the fusion occurred within intronic DNA and the exon structure was preserved. If it did, we checked whether the fusions were in-frame, using the most common frame of the gene's isoforms. Finally, we grouped candidates that predicted the same genomic breakpoint, aggregated read counts and selected the sequence with the most spanning reads as a representative. For each candidate that was identified by JAFFA we use the spanning reads, spanning pairs, whether the transcriptional breakpoint aligned with exon boundaries and genomic gap to classify then rank the candidates.

Combine multi-sample results

The pipeline described above was executed in parallel for each sample in a dataset. As a final step, we merged the results from all samples, outputting a table of results and candidate fusion sequences.

Reference data

The reference transcriptome sequences (GENCODE version 19), exon structure information and human genome version hg19 were downloaded from UCSC. The reference transcriptomic data are provided with the JAFFA package.

Datasets used to assess JAFFA

JAFFA's sensitivity and false discovery rate were evaluated on three simulated datasets and four RNA-seq datasets from cancer cell lines and primary patient samples. Together, these datasets span a range of read length from 50 bp up to full-length transcripts.

First, we used simulated data provided by FusionMap [8,32] to assess JAFFA's power. The FusionMap dataset consisted of 57,000 75 bp pair-end RNA-Seq reads. Fifty fusion events were simulated, with a range of coverage levels. However, background reads from non-fusion genes were absent. Therefore we simulated a second dataset to validate JAFFA's false discovery rate by generating 20 million 100 bp paired-end RNA-Seq reads without fusion events - the BEERS dataset. The simulation was performed using BEERS [33] with default parameters. Finally, a third dataset was simulated to assess JAFFA and alternative tools on long paired-end reads of 250 bp, similar to those expected by the MiSeq platform - the MiSeq dataset, containing 120 fusions. Twenty fusions were simulated at each of 1×, 2×, 5×, 10×, 50× and 100× average coverage across the fusion gene. Fusions were created by randomly selecting two coding transcripts from the RefSeq annotation, randomly selecting an exon edge as a breakpoint, and joining the start sequence of one transcript with the end of the other. Reads from fusion gene were generated with a MiSeq error profile using ART [34] and combined with a BEERS simulation of 5 million read-pairs from non-fusion genes. The fragment length was set to 500 bp with a standard deviation of 100 bp. The BEERS and MiSeq datasets are available from the JAFFA website.

Next, we assessed JAFFA's performance using RNA-Seq of several breast cancer cell lines, for which numerous fusions have previously been reported and validated. We did this for a range of read lengths: first, we ran the Assembly mode on 50 bp paired-end reads from Edgren *et al.* [22]. The Edgren dataset contained between 14 and 42 million, 50 bp paired-end reads of each of the BT-474, SK-BR-3, KPL-4 and MCF-7 cell lines (SRA accession SRP003186). Next we used the ENCODE dataset containing 40 million 100 bp paired-end reads of the MCF-7 cell line (SRA accession SRR534293) to assess JAFFA's Direct mode [21]. We also assessed the Direct mode on an MCF-7 transcriptional profiling dataset provided by PacBio [20]. The PacBio dataset consisted of 44,531 non-redundant consensus sequences. In the BT-474, SK-BR-3, KPL-4 and MCF-7 cell lines, used in the Edgren dataset, a total of 99 fusions have previously been validated (Additional file 3) [22,35-38]. We used these fusions as our set of true positives. It is worth noting that not all previously published fusions are identified in all datasets. This is likely not only because of limitations by fusion detection tools, but also because of differences in sequencing methodology, depth and because of variation in cell line preparations from different laboratories. The concordance between different datasets of the MCF-7 cell line is provided in Additional file 1: Figure S4.

Finally, we ran JAFFA on 100 bp paired-end RNA-Seq from a large glioma study (SRA accession SRP027383)

[23]. From the full dataset of 272 samples, we selected a subset of 13 samples to form our glioma validation dataset (Additional file 4). Each of these samples contained two or more validated in frame fusions, with 31 true positives in total (Additional file 4).

Comparison against competing tools

We compared JAFFA to four of the most widely used fusion detection methods; TopHat-Fusion 2.0.13 [6], SOAPfuse 1.26 [39], DeFuse 0.6.2 [7] and FusionCatcher 0.99.3d [40]. This choice was based on the results from several studies [6,9,10,22], along with our own assessment of a broader selection of tools using the Edgren and FusionMap datasets (summarised in Additional file 1: Table S1). TopHat-Fusion and DeFuse are older fusion finding programs, but are used broadly. FusionCatcher and SOAPfuse have been released more recently and promise superior performance over existing tools. Running parameters for each tool can be found in Additional file 1: Methods 2 and a shell script to reproduce the results from JAFFA is provided as Additional file 5. For the analysis of sensitivity and specificity, we only counted fusion gene pairs with multiple breakpoints once. True positives were identified by their gene name. Although JAFFA reports fusion names in order of fusion orientations, any order of gene names was accepted and different gene aliases were also considered.

Results and discussion

JAFFA shows good sensitivity and a low false discovery rate on simulated data

The performance of JAFFA was first assessed using the 75 bp paired-end reads of the FusionMap simulation. JAFFA was run using all three modes: Assembly; Direct; and Hybrid (Table 1). JAFFA's Assembly mode reported 39 out of 50 true positives (78% sensitivity). For the Direct mode this value was lower, at 34 (68% sensitivity). Finally, the Hybrid approach reported more true positives than any other tool (44 out of 50, 88% sensitivity), indicating that even with reads as short as 75 bp, searching for fusions among reads in addition to assembly, improves sensitivity. For all JAFFA modes, true positives were reported as either 'High Confidence' or 'Medium Confidence'. The majority of missed true positives had low read coverage. In contrast to the previous finding of a high false positive rate with the FusionMap dataset (Carrara *et al.* [10,11], Additional file 1: Table S1A), we found that JAFFA, TopHat-Fusion, FusionCatcher, SOAPfuse and deFuse all had very high specificity, with only SOAPfuse reporting one false positive (Table 1).

Because the FusionMap simulation contained no background reads, we assessed JAFFA's false positive rate further with a simulation containing no fusions, but with transcriptional run-through events, the BEERS dataset.

Table 1 A comparison of fusion detection performance on simulated RNA-Seq, 75 bp simulation of 50 fusions

Tool	True positives	Sensitivity	False positives
JAFFA - Hybrid	44 (32/12/0)	88%	0
JAFFA - Assembly	39 (28/11/0)	78%	0
SOAPfuse	37	74%	1
JAFFA - Direct	34 (32/2/0)	68%	0
deFuse	34	68%	0
TopHat-Fusion	27	54%	0
FusionCatcher	Unable to run on a low number of reads		

We ran all three modes of JAFFA in addition to SOAPfuse, TopHat-Fusion, deFuse and FusionCatcher on a simulation set of 57,000 75 bp RNA-Seq read pairs provided with FusionMap. JAFFA had the highest sensitivity when run in Hybrid mode, identifying 44 out of 50 possible fusion events. For all JAFFA modes, no false positives were reported. In parenthesis we show the value at each of JAFFA's classification levels: (high/medium/low) confidence.

On this dataset JAFFA reports no false positives with a rank of 'High Confidence' or 'Medium Confidence' in all modes. However, the Assembly and Hybrid modes reported 23 'Low Confidence' false positives. These false positives were misassembled because of sequence homology along with sequencing errors, SNPs and indels. However, because exon-exon alignment was not preserved, they were ranked as 'Low Confidence'. The Direct mode, which is the nominal mode for the BEERS 100 bp reads, reported a single 'Low Confidence' false positive. Across all datasets we tested, JAFFA almost always classified true positives as either 'High Confidence' or 'Medium Confidence'. Therefore, in practice, we advise that 'Low Confidence' candidates be rejected, unless there is other independent information to support them, such as presence in the Mitelman database. TopHat-Fusion reported two false positives on the BEERS dataset. SOAPfuse reported 111 candidate fusions and FusionCatcher 79; however, in both cases, the tools flagged these false positives as transcriptional run-through events. DeFuse reported 212 false positives, of which 153 were classified as run-through transcription.

JAFFA also demonstrated excellent sensitivity and a very low false discovery rate on 250 bp reads simulating a MiSeq dataset (Table 2). JAFFA reported the highest number of true positives, 86 out of 120, with a sensitivity of 100%, 95%, 85%, 90%, 40% and 20% corresponding to an average fusion gene coverage of 100×, 50×, 10×, 5×, 2× and 1×, respectively. Seventy-one of the true positives were classed as 'High Confidence'. SOAPfuse was the next most sensitive tool with only 61 true positives. JAFFA reported just three false positives. In each case, the spanning read identified came from a true positive, but due to sequence homology with another gene, the gene name and location of the breakpoint in the genome were wrong. This was primarily within a gene family (2 out of 3 false positives), for example, GPR89A-BACE1,

Table 2 A comparison of fusion detection performance on simulated RNA-Seq, 250 bp simulation of 120 fusions

Tool	True positives							False positives
	Total	Fusion gene coverage						
		100×	50×	10×	5×	2×	1×	
JAFFA	86 (71/14/1)	20	19	17	18	8	4	3 (1/1/1)
SOAPfuse	61	16	17	14	11	1	2	79
deFuse	58	20	15	13	10	0	0	118
FusionCatcher	56	18	16	11	9	1	1	78
TopHat-Fusion	22	11	10	1	0	0	0	0

JAFFA and alternative tools were run on a MiSeq simulation of 120 fusions: 20 fusions at each of 1×, 2×, 5×, 10×, 50× and 100× coverage. The simulation included 5 million 250 bp paired-end background reads. In parenthesis we show the value at each of JAFFA's classifications levels: (high/medium/low) confidence.

was reported as GPR89B-BACE1. In all three cases, the corresponding true positive was also reported. SOAPfuse and FusionCatcher each reported almost 80 false positives, all of which were marked as either run-through transcription or events over a short genomic distance. DeFuse reported over 100 false positives, also primarily run-through transcripts. TopHat-Fusion did not report any false positives, but was the least sensitive tool.

JAFFA has excellent performance across a range of read lengths on cancer RNA sequencing

Short reads (50 bp)

On the Edgren dataset, SOAPfuse reported the highest number of true positives, 41, with other tools reporting between 27 and 35 (Additional file 1: Table S2A). Of the 40 validated fusions previously published for the Edgren dataset [22,38], 37 were rediscovered by at least one of the tools tested. In addition, eight fusions that had been validated in other datasets [35-37] of the same cell lines were reported by at least one tool. Of the total 48 true positives, JAFFA missed 20, predominantly as a result of failing to be assembled (for example, Additional file 2).

In addition to the true positives, all tools reported a number of additional candidates. A subset of these are likely to be novel true positives, and we attempted to distinguish these from other reported candidates using either of the following criteria: (1) candidates reported by three or more tools, after excluding those marked as run-through transcription (Additional file 1: Figure S5); or (2) candidates where one of the partner genes is a so-called 'promiscuous fusion gene partners', defined as a gene implicated in multiple true positive fusions within the same sample. For example, an unconfirmed candidate, SULF2-ZNF217 was identified by JAFFA in the MCF-7 cell lines. Because MCF-7 harbours multiple validated fusions involving SULF2 (SULF2 partnered with ARFGF2, NCOA3 and PRICKLE2), SULF2-ZNF217 was counted as a probable true positive (Additional file 1: Table S2A). Promiscuous fusion gene partners were also observed to occur within the same sample (the

MCF-7 and BT-474 cell lines) by Kangaspeska *et al.* [38]. Kangaspeska *et al.* noted that some promiscuous fusion gene partners were amplified and speculate the mechanism for multi-fusion formation may involve breakage-fusions-bridge cycles where the breakage repeatedly occurs within the same gene.

The number of reported fusions that were neither true positives, nor probable true positives varied substantially between each tool, from 4 (FusionCatcher) to 221 (TopHat-Fusion). A high number of reported fusions that are not true positives likely indicates a high number of false positives. However, the absolute number of reported fusions is often not as informative as assessing the ranking of true positives, which we did using an ROC style plot (Figure 2A). DeFuse and TopHat-Fusion each provided a probability value to rank candidates on. For other tools, we ranked using the output information that maximised the area under the ROC curve. For both FusionCatcher and SOAPfuse this was the number of spanning reads. Probable true positives were excluded from the plot. SOAPfuse, FusionCatcher and JAFFA ranked most known fusions high, however SOAPfuse achieved far greater sensitivity than all other tools without compromising on specificity.

All tools had similar computational performance, with the exception of TopHat-Fusion taking longer to run (27 h on a single core of a modern computing cluster compared to under 11 h for all others). Unlike the other tools, JAFFA's RAM utilisation in assembly mode was not constant, but scaled with the input reads due to the *de novo* assembly (Additional file 1: Figure S6A and B).

Long reads (100 bp)

JAFFA's Direct mode, which is suitable for reads of 100 bp and longer was assessed on the ENCODE MCF-7 data (Figure 2B, Additional file 1: Table S2B). JAFFA reported the highest number of true positives (27) of the fusion detection tools and a large number of probable true positives (6), however JAFFA also reported the highest number of other detections (114). These were largely classified as 'Medium Confidence' (91% of candidates) and

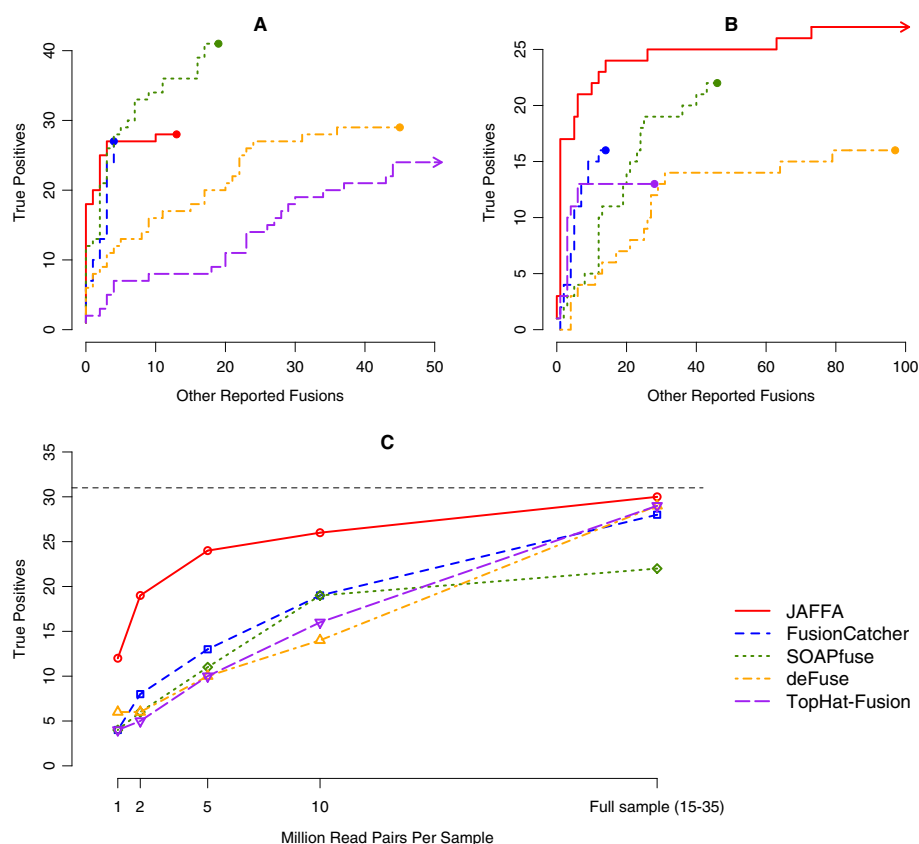


Figure 2 Performance of JAFFA and four other tools on cancer RNA-Seq. **(A)** A ROC-style curve for the ranking of candidate fusions in the Edgren dataset. The Edgren dataset consists of between 7 and 21 million 50 bp read pairs of the BT-474, SK-BR-3, KPL-4 and MCF-7 cell lines. The number of true positives fusions are plotted against the number of other reported fusions from a ranked list of fusion candidates. Probable true positives (see text for detail) are removed. Higher curves indicate a better ranking of the true positives. For each fusion detection tool, we ranked the candidates using the tools own scoring system, or if absent, the supporting data that maximised the area under the curve. SOAPfuse ranked true positives higher than other tools, followed by FusionCatcher and JAFFA. **(B)** On long read data - the ENCODE dataset consisting of 20 million 100 bp read pairs of the MCF-7 cell line - JAFFA ranks true positives higher than any other tool. **(C)** JAFFA's sensitivity is confirmed on a second long read dataset - 13 glioma samples with read depths in the range of 15 to 35 million 100 bp read-pairs. JAFFA identifies 30 of the 31 true positives (total true positives are indicated by the dashed line). Downsampling the data to mimic smaller read depths indicates that JAFFA has similar sensitivity with 2 million read pairs per samples as other tools on 10 million read pairs per sample.

supported by only a single read (89%) (Additional file 1: Table S2B). Thirty-three percent of the other reported fusions were intrachromosomal, and 21% had a genomic gap of less than 3 Mb. Many involved a non-linear ordering of genes. The proportion of local rearrangements were consistent with fusions in the Mitelman database [17] (Additional file 1: Figure S3). We note that JAFFA's Direct mode reported very few false positives on the simulated datasets, and only a single false positive was marked as 'Medium Confidence', this is in contrast to the characteristics of the other reported fusions from real data and may indicate that at least some have a biological origin. Furthermore, they are unlikely to be artifacts from reverse transcriptase template switching during library preparation [41,42], because the breakpoints align with exon-boundaries, suggesting that the fusion event occurred prior to splicing. An interesting possibility, is that the

unknown positives are rare trans-splicing events, such as those found in normal tissue [43,44]. These are also often localised [45,46]. Despite the larger number of unvalidated detections, JAFFA outperformed all other tools in its ability to rank known true positives before other positives (Figure 2B). Again, probable true positives were excluded from the ROC curve. Finally, we compared JAFFA's Direct mode against the Hybrid and Assembly modes (Additional file 1: Figure S7, Table S3), which confirm that there is no advantage in performing an assembly for longer reads (≥ 100 bp). On the contrary, assembly requires substantially more computational resources (Additional file 1: Figure S6C and D).

As a validation of the superior performance of JAFFA with 100 bp reads, we assessed a second dataset consisting of 13 glioma samples with 31 validated fusions. JAFFA detected the highest number of true positives (30

out of 31) and the highest number of probable true positives (45) (Additional file 1: Table S2C). Many of the probable true positives can be explained as out-of-frame fusions. Bao *et al.* identified 147 out-of-frame fusions that were not followed up for validation. TopHat-Fusion and DeFuse reported the equal second highest number of true positives (29), however, we note that the fusions validated by Bao *et al.* were first identified as the intersection of candidates reported by these two tools, so it is expected that they should have close to perfect recall. In an attempt to avoid this bias that favours TopHat-Fusion and DeFuse we next downsampled the dataset to depths of 1, 2, 5 and 10 million read pairs per sample. Across the range of read depths, JAFFA had significantly higher sensitivity in all cases (Figure 2C, Additional file 1: Table S4), while consistently ranking those true positives highly (Additional file 1: Figure S8). For example, we found that with just 2 million read pairs JAFFA achieved the same sensitivity as all other tools achieved on 10 million read pairs, without compromising the false discovery rate (Additional file 1: Figure S9). The sensitivity of JAFFA comes from its ability to reliably call fusions with very low coverage. For example, three of the true positives detected exclusively by JAFFA on the 2 million pair dataset, had just a single read supporting them. This high sensitivity may allow fusions to be identified in samples with low tumour purity or in samples in which a particular fusion is only present in a clone which is a small proportion of tumour cells.

The other positives reported by JAFFA on the full depth dataset, of which there were approximately 300 per sample, displayed similar characteristic to those in the ENCODE dataset, such as a high number of localised rearrangements (Additional file 1: Figure S3). These were primarily supported by a single read (80%) (Additional file 1: Table S2) and aligned with exon-boundaries. If needed, these calls could be removed by requiring multi-read support. We applied a multi-read requirement to the Edgren, ENCODE and gliomas datasets and found that the sensitivity was only slightly reduced (Additional file 1: Figure S10, Additional file 1: Table S2).

On 100 bp reads, all tools were comparable in terms of computational performance (Additional file 1: Figure S6E and F). On the ENCODE dataset, containing 20 million read-pairs, the fusion finding programs took from 7 to 20 h on a single core and 6 to 13 GB of memory. JAFFA required 16 h and 8 GB of RAM. On the gliomas dataset, 13 samples in the range of 15 to 35 million read-pairs were run in parallel. The fusion finding tools required 13 to 50 h and 6 to 13 GB of RAM. JAFFA took 23 h and 11 GB of RAM. Across the Edgren, ENCODE and gliomas datasets, FusionCatcher was consistently the fastest and SOAPfuse consistently used the least memory.

Ultra-long reads and pre-assembled transcriptomes

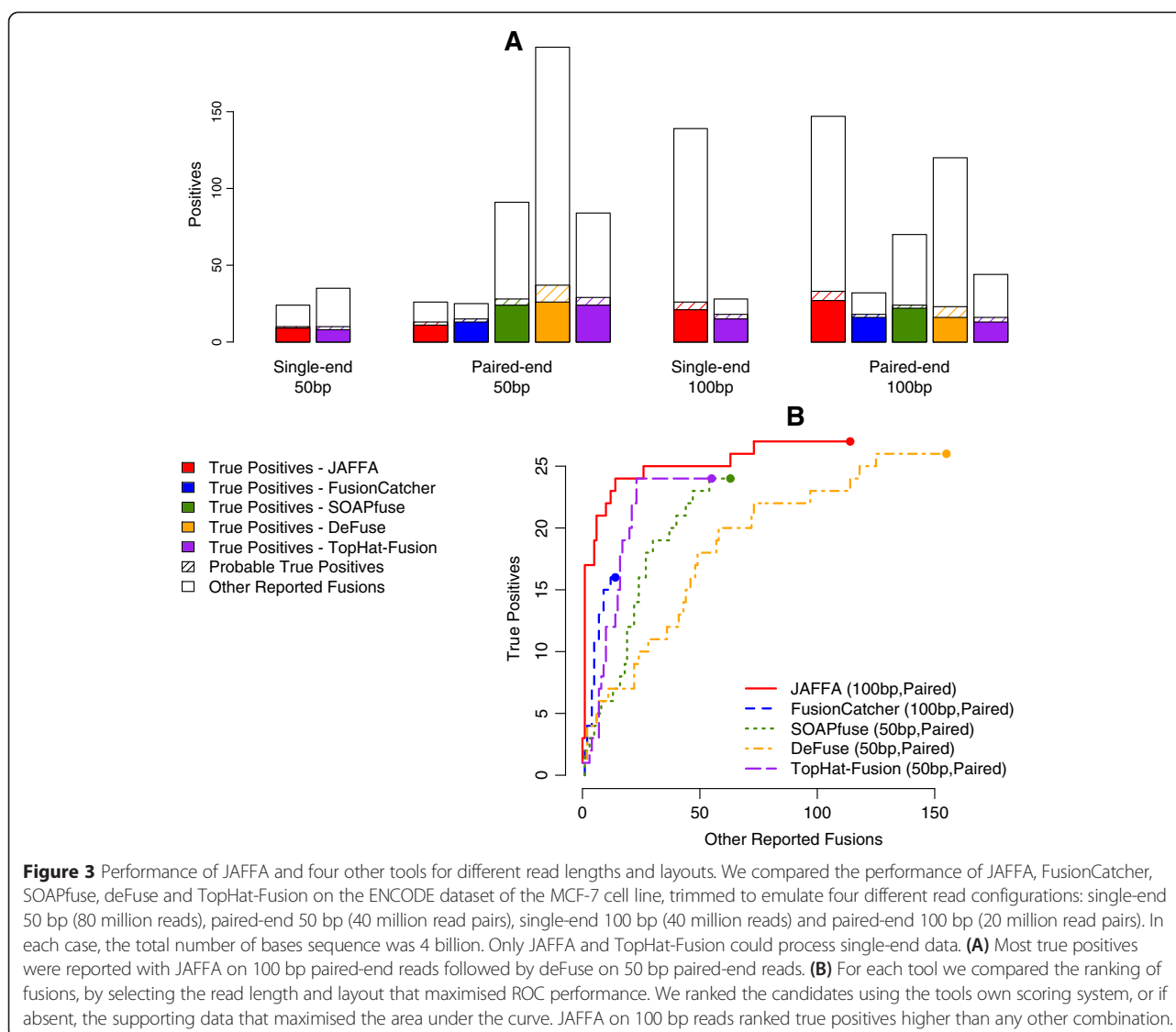
Read lengths are increasing, and technologies such as Ion Torrent, MiSeq and PacBio can already produce reads from several hundred bases up to several kilobases. JAFFA is intrinsically designed for the analysis of such data, because it is based on the idea of comparing transcriptomes. By contrast, it is unclear how well other short read tools work on these data. For example, SOAPfuse, FusionCatcher and deFuse require paired-end reads. TopHat-Fusion cannot be run on ultra-long reads with its recommended aligner, bowtie, because bowtie only aligns reads 1,024 bp and shorter.

To assess the performance of fusion detection methods on ultra-long read data we used the PacBio MCF-7 dataset that has an average sequence length of 1,929 bp. JAFFA was run using the Direct mode and compared with PacBio's own fusion predictions, released with the data [20] (software unavailable). We were unable to successfully run Bowtie2, which aligns longer reads, with TopHat-Fusion. JAFFA reported a similar number of true or probable true positives as the PacBio method (17 compared to 16), but fewer other positives (5 compared to 66). The five unknown positives reported by JAFFA, were also predicted by PacBio. One of these was also predicted by JAFFA in the ENCODE dataset. These results indicate that JAFFA has excellent specificity on ultra-long reads, while still achieving sensitivity similar to tools purpose built for such reads.

Optimal choice of read layout and length

Using the ENCODE dataset, we next addressed the questions of whether paired-end reads perform better than single-end reads, and whether there is any advantage in using 100 bp reads over 50 bp. This question aims to inform experimental design when the sequencing costs of 100 bp, 50 bp, single-end and paired-end are similar for a given number of total bases sequenced. The ENCODE dataset has 100 bp paired-end reads, and was used to create pseudo single-end reads, by selecting one read from each pair, and pseudo 50 bp reads, by trimming off the final 50 bases of each read using FASTX-Toolkit [47]. JAFFA's Assembly mode was run on the 50 bp reads and the Direct mode was run on the 100 bp reads. Each dataset was created with 4 billion sequenced bases, that is, 20 million 100 bp pairs, 40 million 100 bp single-end reads, 40 million 50 bp pairs and 80 million 50 bp single-end reads. Note that the 20 million 100 bp pairs were the same dataset used for the 100 bp validation presented earlier in this manuscript.

When considering each combination of read layout, length and fusion finding algorithm, we found that JAFFA with 100 bp paired-end reads produced the highest number of true positives, with a total of 27 (Figure 3A). However, deFuse, SOAPfuse and TopHat-Fusion reported



a similar number of true positives on 50 bp paired-end reads with 26, 24 and 24, respectively. To determine if these tools were effective at separating the true positives from other predictions, we used a ROC-style curve (Figure 3B). For each tool we show the combination of read length and layout that maximised the ROC performance. For SOAPfuse, deFuse and TopHat-Fusion, this was 50 bp paired-end reads and for JAFFA and FusionCatcher, 100 bp paired-end reads. JAFFA on 100 bp paired-end reads not only reported the highest number of true positives, but provided the best ranking of those true positives (Figure 3B). This trend held across a range of sequencing depths (250 million and 1 billion sequenced bases, Additional file 1: Figures S11 and S12).

Taken together with the results from the simulation, Edgren and glioma datasets, we recommend that datasets with 50 bp paired-end reads be analysed with

SOAPfuse. However any datasets with reads longer than 50 bp or single-end reads should be analysed with JAFFA. When considering how to design an experiment to detect fusion genes, it appears that optimal performance is obtained with 100 bp paired-end sequencing followed by analysis using JAFFA.

Conclusions

We have presented JAFFA, a method for the discovery of fusion genes in cancer transcriptomes by comparing them to a reference transcriptome. The cancer transcriptome is either a set of contigs created by *de novo* assembly of short reads or the reads themselves for longer read sequencing. Therefore one major advantage of JAFFA over previous methods is that it detects fusions using RNA-seq reads of any length, with either single or paired-end reads. JAFFA also provides a simple and

effective method of ranking fusions based on read support and exon-exon boundary alignment. This approach means that we avoid restrictive filtering that may reduce sensitivity.

A limitation of our approach is that JAFFA is not sensitive to fusion genes incorporating intronic or intragenic sequence, because the reference includes only exonic sequence. Moreover, JAFFA down ranks fusions when the breakpoint occurs within an exon, rather than at the boundary. In this case the fusion is ranked as 'Low Confidence'. These two classes of fusions are rare [48,49] and we argue that on balance, the overall improvement in sensitivity and ranking outweighs the potential for these fusion types to be missed. In addition, because JAFFA reports whether a fusion is found in the Mitelman database, fusions classified as 'Low Confidence' that are recurrent in cancer remain identifiable to the user.

In nearly all scenarios we tested, JAFFA outperformed other methods for identifying fusions. The only exception was on 50 bp paired-end reads, where SOAPfuse had the best performance. When we examined the optimal sequencing read layout and length for fusion detection, we found that JAFFA was the most sensitive on 100 bp pair-end reads compared with any other scenario or tool.

The pipeline we have presented is customisable, such that component programmes, for example, the assembler or aligner, can be easily swapped to current state-of-the-art software. Known fusions that were missed by JAFFA on 50 bp reads were lost during the assembly stage. Transcriptome assembly is still maturing, hence there is potential for JAFFA to produce even better fusion detection sensitivity on short reads in the future.

The validation of JAFFA on simulation and RNA sequencing of cancer revealed that our approach has excellent power. In comparison to other fusion detection methods, we found in every scenario with reads longer than 50 bp, JAFFA had the best ranking of true positives above other detections. This included standard short read sequencing, contemporary longer read lengths such as MiSeq 250 bp and ultra-long read PacBio sequencing. This makes JAFFA a fusion detection method that can accommodate the fast pace of change in sequencing technologies.

Additional files

Additional file 1: Includes 12 supporting figures and four supporting tables. A description of each is given within the file.

Additional file 2: Performance of four transcriptome assemblers on the Edgren dataset. A table of which true positive breakpoint sequences were assembled by Trinity, Oases, TransABySS and SOAPdenovo-Trans on the Edgren dataset. Oases assembled the highest number of true positive breakpoints with 31.

Additional file 3: Fusion genes in the BT-474, SK-BR-3, KPL-4 and MCF-7 cell lines. A list of the true positive fusion genes used in the validation of JAFFA on the Edgren and ENCODE dataset, along with a list of the probable true positives, and the fusion calls from JAFFA, FusionCatcher, SOAPfuse, defuse and TopHat-Fusion.

Additional file 4: Fusion genes in the glioma dataset. A list of the true positive fusion genes, probable true positives and results from JAFFA, SOAPfuse, defuse and TopHat-Fusion for the gliomas dataset.

Additional file 5: JAFFA commands. This script provides commands to reproduce the results from JAFFA and other tools shown in the manuscript.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NMD, IJM and AO conceived of ideas in this manuscript. NMD wrote all the code and performed all the analysis. NMD, IJM and AO wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Simon Sadedin for his help with Bpize, Aliaksei Holik and Alex Gout for their feedback as pilot users of the software, Katrina Bell and Maria Doyle for their feedback on this manuscript, and Lorenza Mittempergher, Chong Sun, Astrid Bosma and Paul Ekert for suggestions on the output of JAFFA. This work was made possible through Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRISS. This work was supported by the National Health and Medical Research Council, Australia (Career Development Fellowship 1051481 to AO, Project grant 1051402 to AO, Program Grant 1016647 to IJM, and Fellowship 575581 to IJM).

Author details

¹Murdoch Childrens Research Institute, Royal Children's Hospital, Flemington Road, Parkville, Victoria 3052, Australia. ²Division of Cancer and Haematology, The Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria 3052, Australia. ³Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia. ⁴Department of Genetics, The University of Melbourne, Parkville, Victoria 3010, Australia.

Received: 31 March 2015 Accepted: 21 April 2015

Published online: 11 May 2015

References

- Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*. 2007;7:233–45.
- Edwards PAW. Fusion genes and chromosome translocations in the common epithelial cancers. *J Pathol*. 2010;220:244–54.
- Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009;458:97–101.
- Zhao Q, Caballero OL, Levy S, Stevenson BJ, Iseli C, de Souza SJ, et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci U S A*. 2009;106:1886–91.
- Wang Q, Xia J, Jia P, Pao W, Zhao Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform*. 2013;14:506–19.
- Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*. 2011;12:R72.
- McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*. 2011;7:e1001138.
- Ge H, Liu K, Juan T, Fang F, Newman M, Hoek W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*. 2011;27:1922–8.
- Liu C, Ma J, Chang CJ, Zhou X. FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC Bioinformatics*. 2013;14:193.
- Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, et al. State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int*. 2013;2013:340620.

11. Carrara M, Beccuti M, Cavallo F, Donatelli S, Lazzarato F, Cordero F, et al. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC Bioinformatics*. 2013;14:52.
12. Beccuti M, Carrara M, Cordero F, Donatelli SCR. The structure of state-of-art gene fusion-finder algorithms. *OA Bioinformatics*. 2013;1:2.
13. Zhao S. Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. *PLoS One*. 2014;9:e101374.
14. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rättsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods*. 2013;10:1185–91.
15. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002;12:656–64.
16. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
17. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. [<http://cgap.nci.nih.gov/Chromosomes/Mitelman>]
18. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*. 2012;28:1525–6.
19. Majewski IJ, Mittempergher L, Davidson NM, Bosma A, Willems SM, Horlings HM, et al. Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. *J Pathol*. 2013;230:270–6.
20. PacBio Blog: Data Release: Human MCF-7 Transcriptome. [<http://blog.pacific-biosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html>]
21. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
22. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*. 2011;12:R6.
23. Bao Z-S, Chen H-M, Yang M-Y, Zhang C-B, Yu K, Ye W-L, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. *Genome Res*. 2014;24:1765–73.
24. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012;28:1086–92.
25. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol*. 2014;15:410.
26. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*. 2013;14:328.
27. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22:1760–74.
28. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res*. 2007;17:746–59.
29. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics*. 2011;4:11.
30. JAFFA Homepage. [<https://github.com/Oshlack/JAFFA/wiki>]
31. BBMap Homepage. [<http://bbmap.sourceforge.net>]
32. FusionMap Homepage. [<http://www.arrayserver.com/wiki/index.php?title=FusionMap>]
33. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*. 2011;27:2518–28.
34. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28:593–4.
35. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A*. 2009;106:12353–8.
36. Sakarya O, Breu H, Radovich M, Chen Y, Wang YN, Barbacioru C, et al. RNA-Seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comput Biol*. 2012;8:e1002464.
37. Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res*. 2011;21:676–87.
38. Kangaspeska S, Hultsch S, Edgren H, Nicorici D, Murumägi A, Kallioniemi O. Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms. *PLoS One*. 2012;7:e48745.
39. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol*. 2013;14:R12.
40. Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014. doi: <http://dx.doi.org/10.1101/011650>.
41. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics*. 2006;88:127–31.
42. Houseley J, Tollervey D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*. 2010;5:e12271.
43. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res*. 2012;22:1231–42.
44. Wu C-S, Yu C-Y, Chuang C-Y, Hsiao M, Kao C-F, Kuo H-C, et al. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res*. 2014;24:25–36.
45. Li X, Zhao L, Jiang H, Wang W. Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J Mol Evol*. 2009;68:56–65.
46. Gingeras TR. Implications of chimaeric non-co-linear transcripts. *Nature*. 2009;461:206–11.
47. FASTX Toolkit. [http://hannonlab.cshl.edu/fastx_toolkit/index.html]
48. Novo FJ, de Mendibil IO, Vizmanos JL. TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*. 2007;8:33.
49. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*. 2008; Chapter 10:Unit 10.11.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

