

SOFTWARE

Open Access



# PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events

Osagie G. Izuogu<sup>1\*</sup> , Abd A. Alhasan<sup>1</sup>, Hani M. Alafghani<sup>2</sup>, Mauro Santibanez-Koref<sup>1</sup>, David J. Elliot<sup>1</sup> and Michael S. Jackson<sup>1</sup>

## Abstract

**Background:** Transcripts, which have been subject to Post-transcriptional exon shuffling (PTES), have an exon order inconsistent with the underlying genomic sequence. These have been identified in a wide variety of tissues and cell types from many eukaryotes, and are now known to be mostly circular, cytoplasmic, and non-coding. Although there is no uniformly ascribed function, several have been shown to be involved in gene regulation. Accurate identification of these transcripts can, however, be difficult due to artefacts from a wide variety of sources.

**Results:** Here, we present a computational method, PTESFinder, to identify these transcripts from high throughput RNAseq data. Uniquely, it systematically excludes potential artefacts emanating from pseudogenes, segmental duplications, and template switching, and outputs both PTES and canonical exon junction counts to facilitate comparative analyses. In comparison with four existing methods, PTESFinder achieves highest specificity and comparable sensitivity at a variety of read depths. PTESFinder also identifies between 13 % and 41.6 % more structures, compared to publicly available methods recently used to identify human circular RNAs.

**Conclusions:** With high sensitivity and specificity, user-adjustable filters that target known sources of false positives, and tailored output to facilitate comparison of transcript levels, PTESFinder will facilitate the discovery and analysis of these poorly understood transcripts.

**Keywords:** RNAseq, circRNA, PTES, Splicing, mRNA processing, Software

## Background

Recently, there has been an increased interest in a novel class of transcripts where the exon order differs from that found in the genome [1–4]. Once considered cloning artefacts [5] or products of aberrant splicing [6], it is now established that the majority of these molecules represent circular RNA species (circRNAs) [2–4, 7], although some linear transcripts have been reported [1, 8, 9]. Thousands of these novel transcripts have now been identified in a variety of eukaryotic cells [3, 10], many are conserved across species [2, 11], suggesting functional relevance, and two (from *CDR1* and *SRY*) have been shown to harbour numerous miRNA binding sites and act as miRNA

sponges to modulate gene expression [4, 12]. Recent reports also implicate circRNAs in synaptic development [11] and some have expression patterns that correlate with diseases [13–15] and may act as biomarkers for ageing [16]. Despite these reports, the function of the vast majority of these transcripts has not been established.

The defining feature of these transcripts at the sequence level is the presence of a splice junction with exons in an order inconsistent with their position in the genome. As this feature alone does not enable inference of structure or mechanistic origins, we use the term Post-Transcriptional Exon Shuffled (PTES) transcripts to collectively describe this population of RNA molecules [1]. Recent reports have shown that the vast majority of these transcripts emanate from known genes [2, 17, 18], utilise known splice junctions, and that their biogenesis competes with splicing

\* Correspondence: g.izuogu@ncl.ac.uk

<sup>1</sup>Institute of Genetic Medicine, Newcastle University, Newcastle Upon Tyne, UK

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

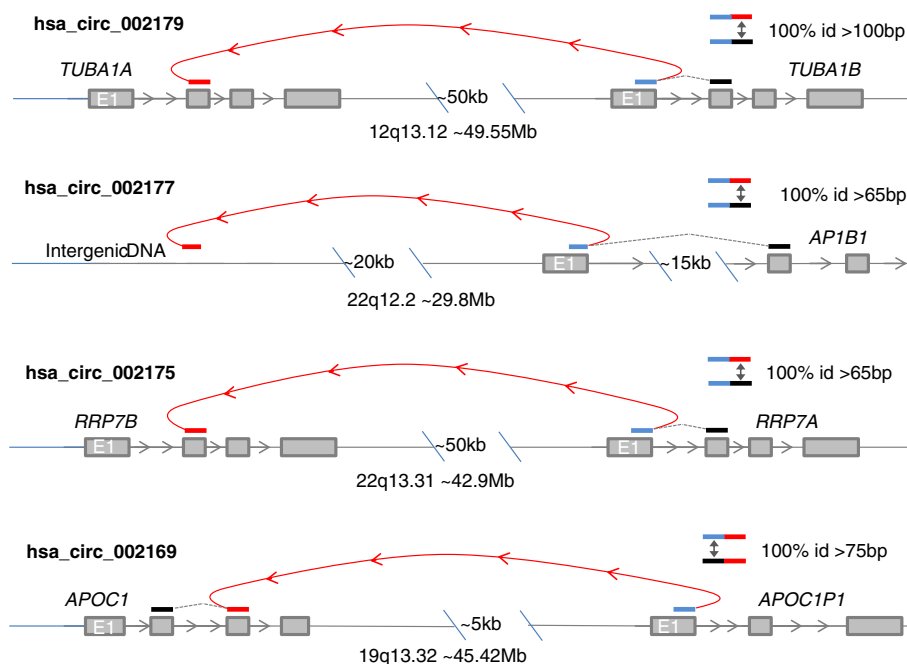
of canonical transcripts [19]. Transcripts arising from PTES specifically exclude chimeric RNAs without known splice junctions, and a class of circular RNAs comprised of spliced introns, ciRNAs [7, 20].

Many computational methods for identifying chimeric RNA molecules from high-throughput RNA sequence data have been described. The majority of these are designed to detect heterotypic trans-splicing and fused genes, so are not suitable for PTES detection [21–27], or require post processing steps to analyse reads supporting different types of splicing events [28, 29]. Recently, however, a number of programs for PTES discovery have been described and used primarily for circRNA characterisation [2–4, 16, 17, 30–32]. Most analyse reads which fail to fully align to a reference sequence, and split these into two or more segments which are then independently aligned to define rearrangements. Some make use of gene annotation to guide discovery [3, 32], while others adopt an unbiased genome-wide approach to capture structures which do not utilise known splice junctions or are non-genic [4, 16, 17, 30, 31]. In addition, the occurrence of PTES can be inferred when two paired end reads map to the transcriptome in a configuration that is not consistent with a linear transcript [3, 16, 30, 32].

The identification of PTES exon junctions within RNAseq data is, however, confounded by known artefacts. False positives can arise from template switching

during cDNA synthesis [1, 5, 9, 33, 34], from genes with duplicated exons [35], from transcription read-through between genes in close proximity due to weak termination signals [36, 37], and from closely related genes within duplicons or tandem arrays [4]. Although experimental enrichment has been combined with informatic approaches to define bona-fide circRNAs [2–4], many classes of false positive structures are not directly excluded by existing identification methods. For example, reads defining 7 of the 20 most abundant human circRNAs reported by Memczak *et al.* [14] map with high sequence identity to the reference sequence and include 4 which are indistinguishable from linear RefSeq entries (Fig. 1). Furthermore, a recent experimental analysis of previously identified PTES transcripts concluded that many are template switching artefacts [38], and template switching predominantly leads to rearrangements where the breaks do not occur at splice junctions [5].

Here we present a method, PTESFinder, that identifies putative PTES structures by mapping RNAseq reads to sequence models generated using existing transcript annotation. It then applies a series of mapping and alignment filters to systematically remove known classes of false positives. It does not make use of paired end (PE) mapping information as the lack of intervening sequence precludes such filtering and may affect specificity. We first describe the implementation of this method, and



**Fig. 1** Examples of Intragenic False Positives. Schematic diagrams showing inferred structure and key sequence relationships for 4 of the 20 most abundant circRNAs reported in [3]. In each case, the inferred structure shares 100 % identity to a linear transcript spanning the defining exon-exon junction. Within the top 20, hsa\_circ\_002174, 002165 and 002164 show similar patterns of identity to multiple genomic locations. Blue – Inferred Donor Exon, Red – Inferred Acceptor Exon, Black – upstream or downstream RefSeq exon sharing 100 % identity to donor or acceptor exon at junction. Approximate chromosomal locations (HG19) are shown

then investigate the effects of different filtering criteria. The program requires certain user adjustable parameters; we therefore also explore systematically the choice of these parameters. Finally, through analysis of real and simulated data, we compare PTESFinder to other publicly available methods [4, 16, 30, 31] which have been used to identify circRNA transcripts in both cell lines and tissues.

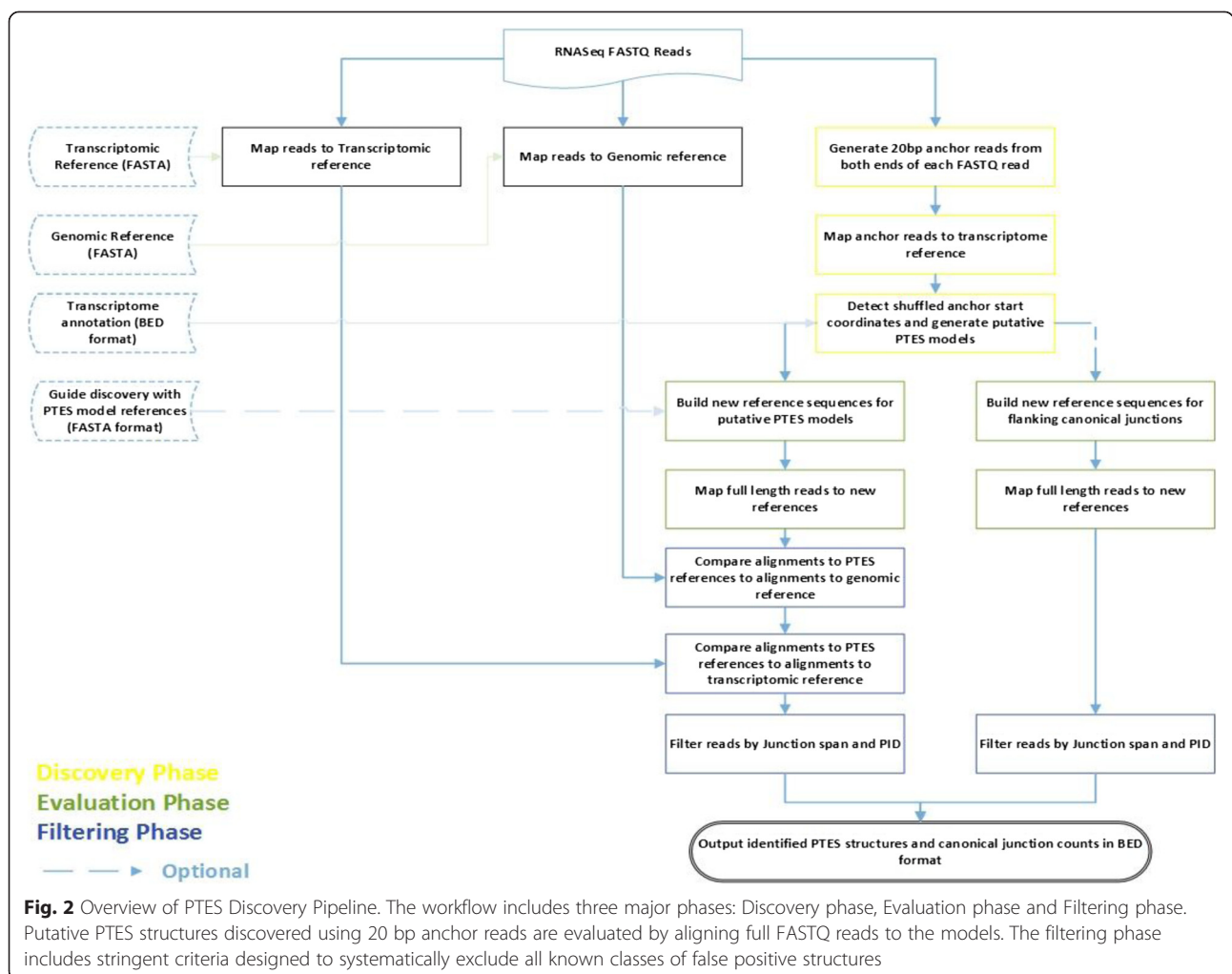
## Implementation

### Pipeline for PTES discovery

PTESFinder requires as input files: RNAseq data in FASTQ format [39], genomic reference in FASTA format, and an annotated transcriptome reference in BED format [40]. The pipeline is split into three phases (Fig. 2): A discovery phase to identify putative PTES structures within RNAseq data and define PTES transcript models, an evaluation phase to assess these PTES models, and a filtering phase to exclude potential false positives.

### Discovery phase

Short sequences from both ends of each FASTQ read (anchors, default length: 20 bp) are aligned to the transcriptome reference using Bowtie [41] with tolerance for a single mismatch. Pairs of anchors from the same read that map to the same gene and in the same orientation, but which map in inverted order with respect to their order in the sequencing read, are then identified. This excludes reads emanating from fused genes and sense-antisense template switching artefacts. Retained anchor pairs are then used to determine the exon junctions which define putative PTES events and create sequence models (constructs) of the inferred products. These constructs are generated by concatenating the last 65 bp of the 5' exon and the first 65 bp of the 3' exon. The segment size of 65 is used by default, with the full exon sequence used if an exon is smaller than 65 bp. This parameter is adjustable to accommodate various RNA-seq read lengths, and we recommend that it be at least 10 bp shorter than the read length to ensure that only reads mapping across PTES defining junctions are



processed in subsequent filtering steps (although as described below these filters would eliminate such reads).

### **Evaluation phase**

All the original reads are then re-mapped to PTES constructs generated in the discovery phase. This serves two purposes. First, as RNAseq reads can be short, this enables reads with putative PTES exon junctions close to their termini to be accurately mapped. Second, it enables read mapping scores obtained using the PTES constructs to be directly compared to scores obtained from genomic and transcriptomic alignments for filtering purposes (see below). Optionally, evaluation can also be 'guided' by supplying constructs of previously discovered PTES structures, effectively bypassing the discovery phase.

### **Filtering phase**

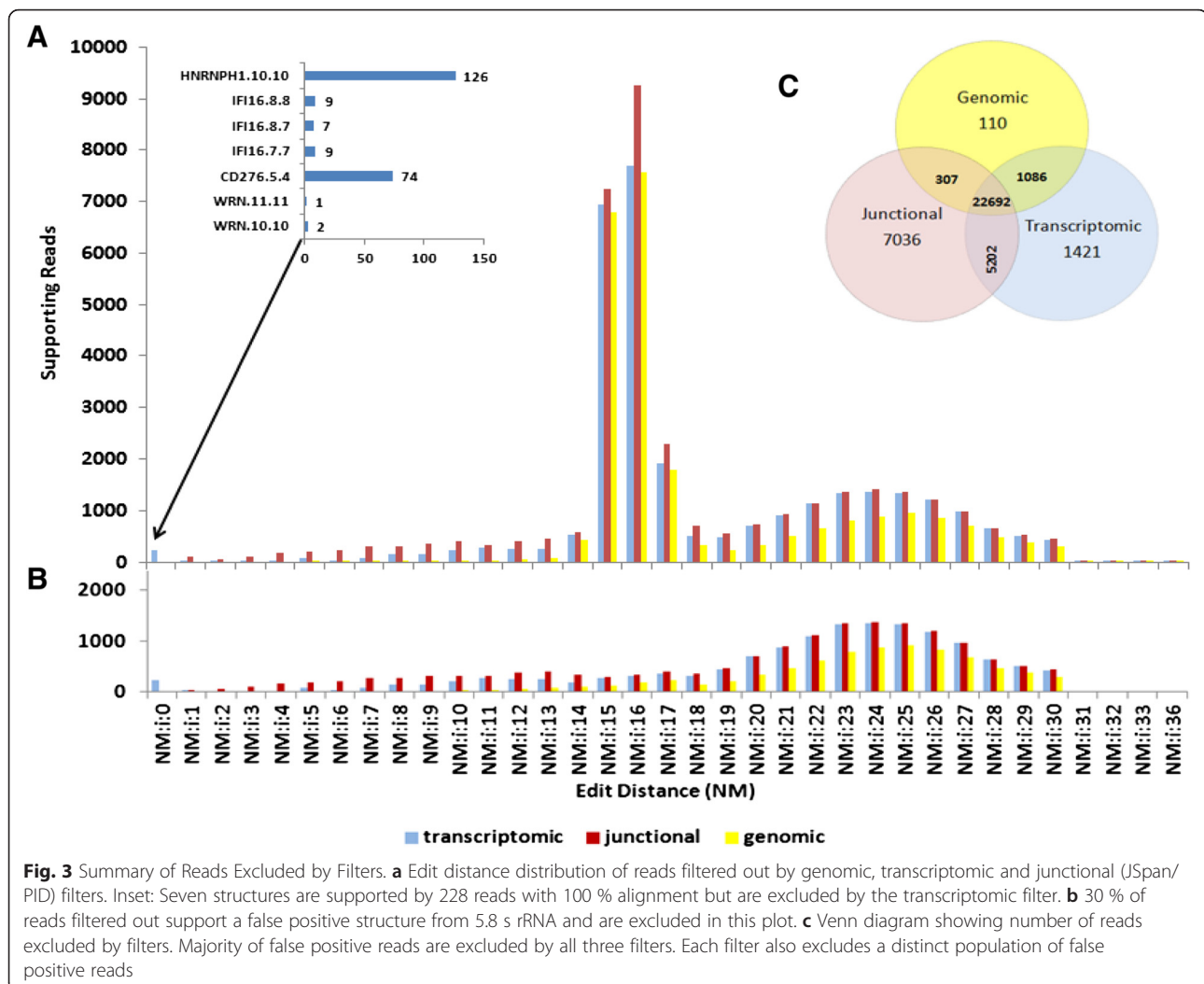
To eliminate potential false positives originating from the genome under investigation, all the original reads are mapped to both genomic and transcriptomic references. The number of edits required for alignment (NM field in SAM format [42]), and the number of perfectly aligned base pairs, are used to remove reads which align as well or better to either of these reference sequences than to the PTES constructs. To reduce template switching artefacts, which have heterogeneous junction points within short regions of often imperfect sequence homology [5], reads which do not align perfectly to the exon junctions which define PTES are also removed using junctional filters. First, a user adjustable minimum junction span (JSpan) parameter is applied to ensure that there are no mismatches or indels within 'n' nucleotides either side of the junction position, where n is an even integer. Second, to eliminate reads with regions of low quality alignment, a user adjustable segment percent identity (PID) parameter is also applied independently to the segments on either side of the PTES junction, such that for a read to be retained both must meet or exceed the specified PID when aligned to the PTES construct. These user adjustable filters rely on alignment summaries provided by the NM field, MD field and Cigar in the SAM files [42]. The output includes the coordinates of the exon end involved in the junctions, a descriptor of the PTES (see Additional file 1 for details) and the number of reads supporting the structure. This is presented in BED format [40]. A second file contains additional information, read counts of all canonical exon junctions from transcripts where a PTES structure has been identified, to facilitate comparison with PTES counts.

### **Assessment of pipeline and comparisons to other methods**

RNAseq data from Jeck *et al.* [2] were analysed at various combinations of JSpan and PID (JSpan range: 4–14; PID range: 60–100 %). All analyses were performed with and without genomic and transcriptomic filters applied to enable reads discarded by each filtering criterion to be identified. The numbers of PTES structures identified and supporting reads were also recorded. To assess sensitivity and specificity, simulated datasets were generated using all published PTES structures within circbase.org [43]. For each dataset, 5000 PTES junctions were randomly selected along with 5000 canonical junctions, and constructs were generated for each junction by concatenating the full sequence of both exons involved in each case. 100 bp simulated reads with random start positions within each construct were then generated. Scripts published by Memczak *et al.* [14] (default parameter values), CIRI v. 1.2 [30] (default parameter values), circRNA\_finder [16] (default parameter values), and MapSplice v. 2.1.5 [31] used in [2] (parameters: --fusion --non-canonical -p16), were compared to PTESFinder by analysing leukocytes cell line RNAseq data (described in [3, 4]), fibroblasts RNAseq data (described in [2]), and simulated data. For each simulation, transcripts correctly identified by each method were determined by comparing genomic coordinates of identified transcripts with the genomic coordinates of transcripts expected to be recovered from within each dataset. The numbers of correctly identified PTES transcripts (true positives – TP), incorrectly identified PTES transcripts (false positives – FP), PTES transcripts incorrectly excluded (false negatives – FN), and canonical junctions correctly excluded (true negatives – TN), were used to estimate sensitivity:  $TP / (TP + FN)$ , specificity:  $TN / (TN + FP)$ , and false discovery rate (FDR):  $FP / (TP + FP)$ .

### **Results and discussion**

PTESFinder uses established RNAseq tools (Bowtie [41], Bowtie2 [44] and Bedtools [45]) to identify putative PTES structures, and then systematically excludes known classes of false positive structures by applying genomic, transcriptomic and junctional (JSpan & PID) filters (see methods). As an initial assessment of PTESFinder function, RNAseq data from human fibroblast total RNA which has previously been mined for circRNAs (sample SRR44975A in [2]), were analysed both with and without the application of the genomic and transcriptomic alignment filters. Reads recovered during analysis, together with alignment



**Fig. 3** Summary of Reads Excluded by Filters. **a** Edit distance distribution of reads filtered out by genomic, transcriptomic and junctional (JSpan/PID) filters. Inset: Seven structures are supported by 228 reads with 100 % alignment but are excluded by the transcriptomic filter. **b** 30 % of reads filtered out support a false positive structure from 5.8 s rRNA and are excluded in this plot. **c** Venn diagram showing number of reads excluded by filters. Majority of false positive reads are excluded by all three filters. Each filter also excludes a distinct population of false positive reads

edit distances of reads identified by each filter applied separately, are shown in Fig. 3a.

#### Filters target overlapping populations of reads but none is redundant

From a total of over 200 million reads analysed, approximately 0.17 % (359837) have shuffled coordinates with respect to exon position (Fig. 3c and Additional file 2: Table S1). However, of these only 44620 (~12.5 %) map to PTES sequence constructs generated during the discovery phase, indicating that most of the reads with rearranged anchor pairs do not map to single genes and/or known exon junctions. Approximately 85 % (37854) of the reads which map to PTES constructs are subsequently removed by the genomic, transcriptomic and junctional (JSpan and PID) filters, with the majority being identified by more than one filter. For instance, over 98 % of reads excluded by the genomic filter are also excluded by the transcriptomic filter, and 60 % (22692) of all filtered reads are identified by all three. Most of these

have high edit distances (>10) indicative of low quality alignment. Despite this, the genomic, transcriptomic and junctional filters (at lowest stringency) uniquely exclude ~0.25 % (110), ~3.2 % (1421) and 15.8 % (7036) of reads mapping to PTES models respectively (Fig. 3a), indicating that none is wholly redundant.

The subset of reads identified specifically by the junctional and transcriptomic filters are defined by low edit distances of between 1 and 10 (Fig. 3a), although a small number of reads excluded by the transcriptome filter (228) map perfectly to putative PTES constructs with NM = 0 (inset). Fig. 3a also reveals a bimodal distribution of mapping qualities for reads excluded by all three filters with peaks at NM = 16 and NM = 24. Upon manual analysis, most of the excluded reads with NM = 16 were found to support a false positive structure from 5.8 s rRNA (NR\_003285.1.1). Comparable rRNA derived structures have been identified previously and filtered manually [4]. In Fig. 3b, reads supporting this



structure have been removed to show the underlying distribution of mapping quality scores.

#### Reads excluded by specific filters have different origins

To investigate the activity of specific filters further, the mapping co-ordinates of reads removed by the genomic filter were first compared to the co-ordinates of annotated pseudogenes and segmental duplications. This established that ~74 % of reads excluded by the genomic filter had superior alignments to segmental duplications, and ~12 % had superior alignments to pseudogenes. The 417 reads identified by the genomic filter but not by the transcriptomic filter were also found to be enriched for reads derived from segmental duplications and pseudogenes (e.g. Additional file 3: Figure S1A).

We next used BLAT [46] to manually investigate the 228 reads excluded specifically by the transcriptome filter which mapped perfectly to putative PTES constructs (NM = 0, Fig. 3b). These support 7 putative PTES structures from 4 genes (Inset, Fig. 3a). However, BLAT analysis established that they all also mapped contiguously with ~100 % identity to the transcriptome due to high sequence identity between neighbouring exons. For example, 126 reads which support a putative single exon PTES structure (exon 10 of *HNRNPH1* circularized) map with ~100 % identity to exons 10 and 11 of the canonical *HNRNPH1* transcript (Additional file 3: Figure S1B) due to high sequence identity between these neighbouring exons. As a result, these reads cannot be taken as supporting evidence for PTES. It is noteworthy that such structures will pass any qualitative filter criterion requiring only unambiguous mapping to PTES constructs, illustrating the value of the transcriptome filter.

Finally, manual analysis of a subset of the 7036 reads identified only by the junctional filters established that these support structures with distinct patterns of sub-optimal mapping, such as low alignment quality specific to only one of the two exons in the structure (e.g. Additional file 3: Figure S1C top 2 panels), and low sequence identity specifically at the junction (e.g. Additional file 3: Figure S1C lower 2 panels), the latter being consistent with the expected pattern of alignment for template switching artefacts [5].

As one further assessment of the filters, we analysed RNAseq data derived from fibroblast RNA which had been pre-digested with RNase R. This selectively removes linear RNAs, and enriches for circRNAs [7, 47], and has been shown to significantly increase the recovery of PTES reads. However, we would anticipate that this would also selectively remove false positives derived from pseudogenes and segmental duplications which mimic PTES structures, without necessarily reducing reverse transcription artefacts such as template switching. Only ~12 % of reads from the RNaseR digested sample

which map to PTES sequence constructs are identified by the genomic and transcriptomic filters (Additional file 2: Table S1), compared to 69 % in the undigested sample. Furthermore, only 17 % of these map to segmental duplications, compared to 74 % in the undigested sample. In contrast, the proportion of reads excluded by the junctional filters is considerably higher after RNaseR digestion, consistent with expectation.

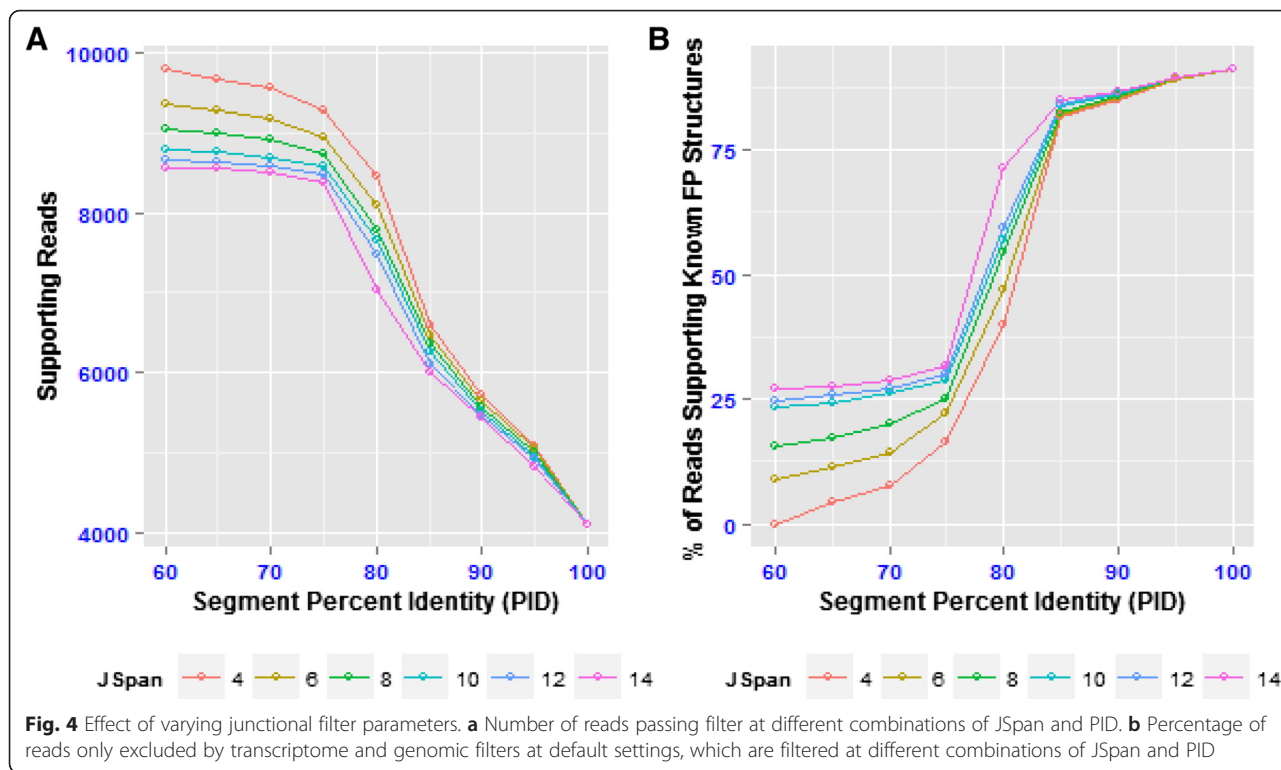
#### PID Has greater impact than JSpan

To investigate the impact of varying the user defined JSpan and PID parameters which comprise the junctional filter, the same data was re-analysed using 54 different combinations of these parameters, both with and without the genomic and transcriptome filters applied (Fig. 3c). This established that varying the PID has a greater impact than varying the JSpan, with 5691 reads filtered with maximal PID (100 %) and lowest JSpan (4) compared to only 1235 reads filtered with the maximal JSpan (14) at lowest PID (60 %). Furthermore, varying the PID between 60 % and 75 % has little impact at any JSpan value, but above 75 % there is a linear relationship with the number of reads filtered. As the default junctional filter parameters failed to identify some reads excluded by the other filters (110 and 1421, Fig. 4a), this analysis was repeated using only these reads to establish the JSpan and PID parameters required to identify them. Over 99 % of these reads are excluded with the most stringent junctional filter parameters (Fig. 4b). Furthermore, the vast majority are filtered with a PID of 85 %, suggesting this is a logical setting for this parameter. The JSpan setting only has a major impact at low PIDs (60–75 %).

#### Specificity, sensitivity and comparison with other methods

To assess the sensitivity and specificity of the pipeline and compare it to other methods, simulated reads were generated from previously identified PTES and associated canonical transcripts, and analysed at various read depths of coverage using default parameters. In addition to assessing PTESFinder for *de novo* PTES discovery, the use of constructs of previously reported structures for guided discovery was also assessed (see methods), as were four publicly available methods which have previously been employed to identify circRNA transcripts: MapSplice v 2.1.5 [31] used in [2], CIRI v. 1.2 [30], circRNA\_finder [16] and the method used by Memczak *et al.* [4].

Results from 100 simulated datasets are presented in Fig. 5a-c, and illustrate that sensitivity varies considerably with coverage, and between methods. At read coverage of 2, the sensitivity of PTESFinder is below 0.6. This can be attributed to PTES junctions occurring

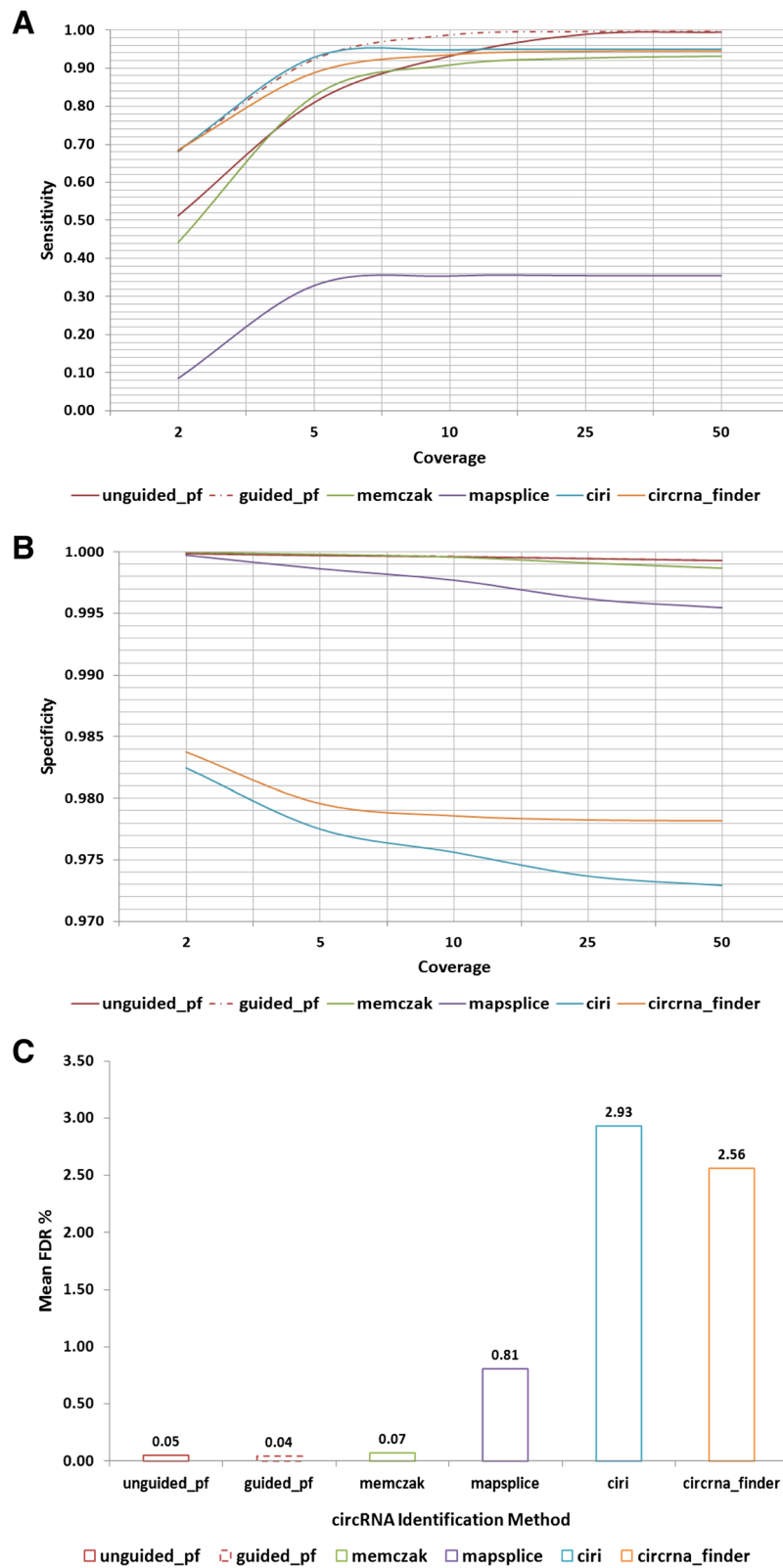


within the terminal 20 bp of reads, as the low tolerance for mismatches during anchor mapping will result in their elimination. However, sensitivity reaches >90 % at coverage of 10 or higher for both guided and unguided analyses, with guided PTESFinder being equally or more sensitive than all other methods at all read depths. Strikingly, the sensitivity of MapSplice is low, remaining below 0.5 at all read depths. In contrast, specificity is over 0.97 for all methods at all read depths (Fig. 5b), although PTESFinder achieves the highest specificities averaged across all depths (over 0.999) for both *de novo* and guided PTES discovery, with all canonical junction reads being correctly identified as such within the simulated data. Only the Memczak method has similar specificity when averaged across all read depths (Fig. 5c).

To compare performance using real data we first re-analysed data from Jeck et al. [2] using all 5 methods (Table 1). To allow direct comparison to PTESFinder, the number of putative circRNA structures identified which utilise 2 RefSeq splice sites was recorded for all other methods (bracketed), as the total numbers include structures from intergenic and intronic regions of the genome. For all 4 samples analysed, CIRI consumed >90Gb of memory, resulting in incomplete analyses. It was therefore not analysed further. Of the remaining 4 methods, PTESFinder identified on average 15 % more structures than the Memczak method and ~70 % more than MapSplice. The latter is consistent with our finding

that MapSplice, which was used in their analysis [2], has low sensitivity at all depths of sequence coverage (Fig. 5a). However, circRNA\_finder reported the highest number of putative circRNA transcripts from both exonic and non-exonic regions of the genome, reporting approximately 31–42 % more structures with RefSeq coordinates than PTESfinder (Table 1).

To investigate the origins of the RefSeq related structures identified exclusively by circRNA\_finder, reads defining these structures from 1 sample (SRR444975) were re-analysed using PTESFinder (Fig. 6a). Of 9287 reads re-analysed, approximately 20 % (1840) are defined as multilocus or sense-antisense fusions, and a further 19 % (1775) are eliminated by the junctional, genomic, and transcriptomic filters indicating likely false positives (Fig. 6b). The remaining 61 % (5672) are not aligned, indicating that their anchors map suboptimally to RefSeq. Furthermore, plotting the distribution of the number of reads supporting each structure identified by circRNA\_finder only, by PTESFinder only, and by both methods (Fig. 6c), revealed that the vast majority of structures identified by circRNA\_finder alone are supported by a single read. This is in sharp contrast to structures identified by both methods, or by PTESFinder alone. While these single-read structures may include *bona fide* low frequency circRNAs, they are also likely to contain false positives caused by suboptimal mapping, consistent with the lower specificity of circRNA\_finder with our simulated data.



**Fig. 5** Sensitivity and Specificity in Comparisons to Other Methods. **a** Sensitivity and **(b)** Specificity of PTESFinder and 4 other publicly available methods (CIRI, circRNA\_finder, MapSplice and Memczak) analysed using simulated data (see methods). **c** Mean false discovery rate % (FDR) of all methods averaged across all read depths analysed



**Table 1** Number of PTES transcripts identified from Human Fibroblast samples using four methods

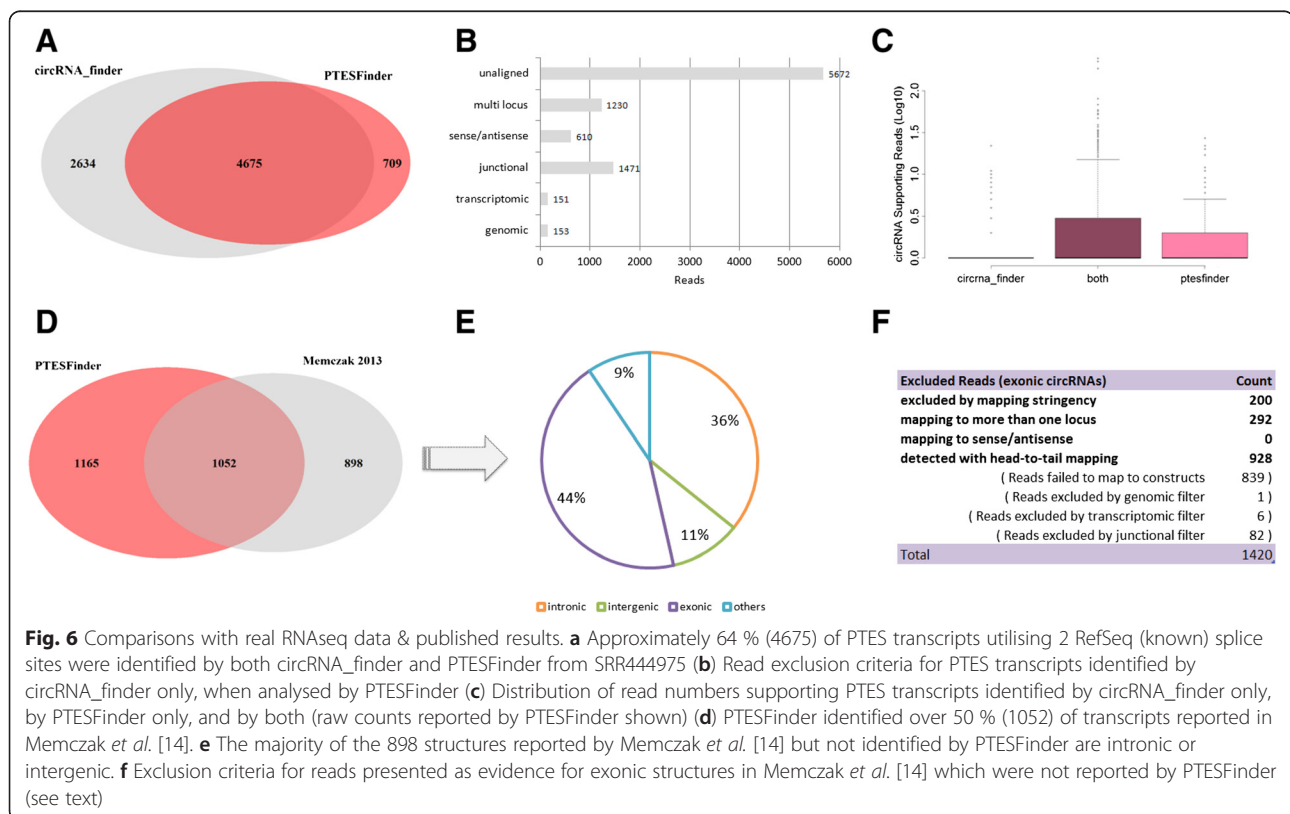
Method		SRR444974	SRR445016	SRR444975	SRR444655
Memczak <sup>a</sup>	Identified	22663 (17752)	22351 (17231)	3733 (2956)	1667 (873)
	Run Time	1993 m	2479 m	2602 m	2061 m
MapSplice <sup>a</sup>	Identified	9701 (7087)	7380 (4891)	2231 (986)	1479 (307)
	Run Time	6167 m	16356 m	7412 m	2605 m
PTESFinder	Identified	25116	24489	5383	2316
	Run Time	1355 m	1963 m	1530 m	1369 m
circRNA_finder <sup>a</sup>	Identified	49901 (32856)	54154 (32186)	11069 (7309)	3130 (2131)
	Run Time	75 m	90 m	80 m	88 m

<sup>a</sup>circRNAs utilizing two RefSeq annotated splice sites in brackets

Runtimes for PTESFinder were also 25–35 % lower than for the Memczak method, and 50–82 % lower than for MapSplice (Table 1), but by far the best runtimes were achieved by circRNA\_finder which utilises the STAR aligner [48]. These were, however, achieved at higher computing memory cost (~30GB).

We then used PTESFinder to analyse RNAseq reads previously mined in two further studies [3, 4]. Consistent with the above, it identified 13 % more distinct structures from leukocyte and HEK293 data than were reported by Memczak *et al.* [4] (2217 as opposed to 1950 Fig. 6d), and 41.6 % more structures than reported by Salzman *et al.* [3] from leukocyte data (1875 as opposed

to 1324, (data not shown)). As both structures and supporting reads were reported by [4], it was possible to re-analyse the 898 structures identified exclusively by their method using PTESFinder. This established that none correspond to structures which PTESFinder is designed to identify (Fig. 6e): 503 (56 %) are derived from intronic, and intergenic regions, and of the 1420 reads supporting the remaining 395 genic structures, 492 were excluded by PTESFinder due to low map quality (200) or multiple map locations (292), 89 reads were excluded by PTESFinder filters, and the remaining 839 possessed at least 1 exon boundary which did not map to known splice junctions (Fig. 6f). Again, while some of these



latter reads will undoubtedly correspond to *bona fide* PTES structures (as a number of genic PTES utilising non-Refseq splice sites have been confirmed experimentally (e.g. [1, 4]), further BLAT analysis established that 13 mapped in a linear fashion to 6 annotated pseudogenes (Additional file 4: Table S2).

Approaches to PTES discovery involve a compromise between the ability to detect all potentially rearranged transcripts, and the ability to identify artefacts generated as a result of the sequence and structural complexity of eukaryotic genomes, and of current library construction methods. It is now clear that the majority of transcripts with re-arranged exon order utilize known exon junctions [2, 18] which are processed by the spliceosome [17, 19]. As a result, methods which utilise existing transcript annotation from the genome under study, such as PTES-Finder and those employed by [3, 32], benefit from the reduced noise inherent in this approach and are suited to quantitative analyses of PTES structures that can be characterized using existing annotations.

The use of known/experimentally verified splice sites does reduce the misidentification of template switching artefacts or unspliced pseudogenes as *bona fide* PTES transcripts. However, it does mean that not all rearranged transcripts will be identified. Although a recent analysis of human data unconstrained by existing annotation suggests that circRNAs which function as miRNA sponges are rare [17], discovery of transcripts which do not utilise known splice sites (including any which are not processed by the spliceosome) requires a genome-wide approach unconstrained by existing annotations. Such approaches are, however, inherently more susceptible to artefacts. The analyses presented above illustrate both the problem of false positive structures, the trade off between sensitivity and specificity in all methods designed to identify rearranged transcripts, and the utility of multiple filters designed to target distinct populations of known artefacts.

## Conclusions

A major challenge in PTES identification is to discriminate between *bona fide* PTES structures and a wide variety of false positives with distinct origins. Currently, no method which has been used for PTES discovery explicitly excludes all known classes of false positive reads. To that end, we have developed PTESFinder to identify both linear and circular PTES transcripts from high throughput RNA-seq data. Compared to publicly available methods recently used in circRNA discovery, PTESFinder achieves higher specificity and sensitivity, and generates output tailored for downstream comparative analyses of transcript abundance, making it an appropriate tool to investigate these RNAs within complex mammalian genomes.

## Availability and requirements

- **Project name:** PTESFinder
- **Project home page:** <http://ptesfinder-v1.sourceforge.net/>
- **Operating system(s):** Linux
- **Programming language:** Shell, Java 1.6
- **Other requirements:** Bowtie 1.1.1 & 2.2.4, BedTools 2.22.0
- **License:** OSI-Approved Open Source (Artistic License 2.0)
- **Any restrictions to use by non-academics:** None

## Additional files

**Additional file 1: Distinct PTES structures identified from dataset mined in Memczak et al. [4].** (XLSX 1443 kb)

**Additional file 2: Table S1.** Analyses of RNASEQ data from human fibroblast cells. (PDF 19 kb)

**Additional file 3: Figure S1.** Example Reads Excluded By Filters.

**A)** Reads filtered out by genomic filter for mapping better to pseudogenes & segment duplicated regions **B)** Reads excluded by the transcriptomic filter for having 100 % alignment to a canonical splice between exons 10 and 11 of HNRNPH1 **C)** Reads excluded by applying the junctional filters, segment PID and JSpan (see text). (PDF 475 kb)

**Additional file 4: Table S2.** circRNA transcripts published by Memczak et al. [4] with 100 % overlap to annotated pseudogenes and excluded by PTESFinder. (XLSX 12 kb)

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

MJ, DE and MSK conceived this project; MSK and OI developed the software; OI, AA and HMA performed *in silico* analyses of PTES transcripts; OI, MSK and MJ wrote the paper; DE provided guidance and manuscript editing. All authors read and approved the manuscript.

## Acknowledgements

This work was supported BBSRC studentship (BB/J014516/1), Leverhulme Trust grant (RPG-2012-795) and Wellcome Trust grant (WT089225MA).

## Author details

<sup>1</sup>Institute of Genetic Medicine, Newcastle University, Newcastle Upon Tyne, UK. <sup>2</sup>Security Forces Hospital, P. O. Box 2748-24268-8541, Makkah, Kingdom of Saudi Arabia.

Received: 11 February 2015 Accepted: 6 January 2016

Published online: 13 January 2016

## References

1. Al-balool HH, Weber D, Liu Y, Wade M, Guleria K, Lang P, et al. Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant. *Genome Res.* 2011;21(11):1788–99.
2. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA.* 2013; 19(2):141–57.
3. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One.* 2012;7(2):e30733.
4. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature.* 2013;495(7441):333–8.
5. Houseley J, Tollervey D. Apparent non-canonical trans-splicing is generated by reverse transcriptase *in vitro*. *PLoS One.* 2010;5(8):e12271.

6. Cocquerelle C, Mascres B, Hétiuin D, Bailleul B. Mis-splicing yields circular RNA molecules. *FASEB J*. 1993;7(1):155–60.
7. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. *Nat Biotechnol*. 2014;32(5):453–61.
8. Dixon RJ, Eperon IC, Hall L, Samani NJ. A genome-wide survey demonstrates widespread non-linear mRNA in expressed sequences from multiple species. *Nucleic Acids Res*. 2005;33(18):5904–13.
9. Wu C-S, Yu C-Y, Chuang C-Y, Hsiao M, Kao CF, Kuo HC, et al. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res*. 2014;24(1):25–36. doi:10.1101/gr.159483.113.
10. Wang PL, Bao Y, Yee M-C, Barrett SP, Hogan GJ, Olsen MN, et al. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One*. 2014;9(3):e90859.
11. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol Cell*. 2015;58(5):870–885.
12. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. *Nature*. 2013;495(7441):384–8.
13. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet*. 2010;6(12):e1001233.
14. Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, et al. Correlation of circular RNA abundance with proliferation—exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Sci Rep*. 2015;5:8057.
15. Li P, Chen S, Chen H, Mo X, Li T, Shao Y, et al. Using circular RNA as a novel type of biomarker in the screening of gastric cancer. *Clin Chim Acta*. 2015;444:132–6.
16. Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, et al. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and Age-dependent neural accumulation. *Cell Rep*. 2014;9(5):1966–80.
17. Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol*. 2014;15(7):409.
18. Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev*. 2014;28(20):2233–2247. doi:10.1101/gad.251926.114.
19. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, et al. CircRNA biogenesis competes with Pre-mRNA splicing. *Mol Cell*. 2014;56(1):55–66.
20. Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, et al. Circular intronic long noncoding RNAs. *Mol Cell*. 2013;51(6):792–806.
21. 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 Suppl 7:S2.
22. Chen K, Wallis JW, Kandath C, Kalicki-veizer JM, Mungall KL, Mungall AJ, et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics*. 2012;28(14):1923–4.
23. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*. 2011;12(8):R72.
24. 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(1):193.
25. Tang S, Riva A. PASTA: splice junction identification from RNA-sequencing data. *BMC Bioinformatics*. 2013;14(1):116.
26. Wu J, Zhang W, Huang S, He Z, Cheng Y, Wang J, et al. SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads. *Bioinformatics*. 2013;29(23):2971–8.
27. Abate F, Acquaviva A, Paciello G, Foti C, Ficarra E, Ferrarini A, et al. Bellerophon: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. *Bioinformatics*. 2012;28(16):2114–21.
28. Philippe N, Salson M, Commes T, Rivals E. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol*. 2013;14(3):R30.
29. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, et al. Open Access A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. 2014.
30. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol*. 2015;16(1):4.
31. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178–8.
32. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-type specific features of circular RNA expression. *PLoS Genet*. 2013;9(9):e1003777.
33. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics*. 2006;88(1):127–31.
34. Odelberg SJ, Weiss RB, Hata A, White R. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res*. 1995;23(11):2049–57.
35. Letunic I, Copley RR, Bork P. Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet*. 2002;11(13):1561–7.
36. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, et al. Transcription-mediated gene fusion in the human genome. *Genome Res*. 2006;16(1):30–6.
37. 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(1):11.
38. Yu C-Y, Liu H-J, Hung L-Y, Kuo H-C, Chuang T-J. Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res*. 2014;42(14):9410–23.
39. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010;38(6):1767–71.
40. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996–1006.
41. 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(3):R25.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
43. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA*. 2014;20(11):1666–70.
44. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012;9(4):357–9.
45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
46. Kent WJ. BLAT — the BLAST -like alignment tool. 2002. p. 656–64.
47. Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res*. 2012;40(7):3131–42.
48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
www.biomedcentral.com/submit

