ASJA: A Program for Assembling Splice Junctions Analysis

Jingjing Zhao a,b,1, Qin Li a,b,1, Yuchen Li a,b, Xiango He a,b, Qiupeng Zheng a,b,*, Shenglin Huang a,b,∗

a Fudan University Shanghai Cancer Center, Key Laboratory of Medical Epigenetics and Metabolism, Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
b Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

Abstract

RNA splicing may generate different kinds of splice junctions, such as linear, back-splice and fusion junctions. Only a limited number of programs are available for detection and quantification of splice junctions. Here, we present Assembling Splice Junctions Analysis (ASJA), a software package that identifies and characterizes all splice junctions from high-throughput RNA sequencing (RNA-seq) data. ASJA processes assembled transcripts and chimeric alignments from the STAR aligner and StringTie assembler. ASJA provides the unique position and normalized expression level of each junction. Annotations and integrative analysis of the junctions enable additional filtering. It is also appropriate for the identification of novel junctions. ASJA is available at https://github.com/HuangLab-Fudan/ASJA.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

RNA splicing is a form of RNA processing in which introns are removed from a newly synthesized precursor messenger RNA producing a mature mRNA containing splice junctions. The human multiexon genes have precise alternative spliced variants, which gives rise to genetic diversity. Aberrations of RNA splicing are associated with a wide range of diseases, such as cancers, neurodegenerative diseases and muscular dystrophies [1,2]. Different from the canonical, normally spliced RNAs, noncanonical splicing processes, resulting in circular RNAs (circRNAs) and fusion gene, can be identified from chimeric RNAs. Fusion genes play a vital role in tumor initiation and progression [3–5]. Chromosomal translocations are the common cause of oncogenic fusion genes. One of the best known examples is BCR-ABL1 that plays important factor in adult acute lymphoblastic leukemia and can be used as a biomarker for chronic myeloid leukemia [6–8]. Anaplastic lymphoma kinase (ALK) gene fusion are present in approximately 5% of non-small-cell lung cancers (NSCLCs) [9], which indicates a new therapeutic target in this molecularly defined subset of NSCLC. Besides, TMPRSS2–ETS fusions can be detected in approximately 50% of localized and approximately 40% of metastatic prostate cancer respectively [10,11]. CircRNAs with back-splice junctions are characterized by single-stranded, covalently closed loop structures. High-throughput sequencing computational approaches have demonstrated that circRNAs are widespread in the transcriptome. The functions of most circRNAs remains largely unexplored, but known functions include binding to microRNAs or proteins [12–14], regulating their parent genes [15,16], and producing protein products [17,18]. In addition, emerging studies have begun to reveal the critical role that some circRNAs play in the nervous system, cancer development and innate immune response [19].

Current tools for RNA-seq analysis perform well for predicting splice sites [20], but software for splice junction detection and quantification is limited. Nellore et al. found a large number of linear exon-exon junctions by aligning approximately twenty-thousand RNA-seq samples from the Sequence Read Archive (SRA) [21]. However, they extracted the splice junctions directly from mapping reads and many of these junctions might have a high false-positive rate due to incorrect read placement, sample-specific variation and regions where the genome is incorrectly assembled. For circRNA detection, there are a handful of tools and database available [22,23]. However, current computational tools for splice junction detection only focus on one or two types of junctions [12,24,25]. To fully evaluate the whole transcriptome based on RNA-seq data, it is necessary to develop a comprehensive tool to identify and characterize different kinds of splice junctions.

In this study, we present a software package called Assembling Splice Junctions Analysis (ASJA) that detects and annotates all splice junctions from RNA-seq data. It processes assembled transcripts and chimeric alignments from the STAR aligner [26] and Stringtie assembler [27]. ASJA provides the unique position and normalized expression level of each junction for comparison. We verified the workflow on a published RNA-seq dataset [28] and showed that ASJA could efficiently
detect splice junctions and was also appropriate for the identification of novel junctions.

2. Methods

2.1. The ASJA Algorithm

ASJA identifies three types of splice junctions, using the following three steps: 1) Alignment and generation of assembled transcripts, 2) Junctions extraction, 3) Junctions annotation and integration.

2.2. Alignment and Generation of Assembled Transcripts

2.2.1. STAR Aligner Settings

FastQC software was used for RNA-seq raw read quality control. The STAR aligner (version 2.5.2a) was used to map all filtered reads with a 2-pass mapping process [29]. In the first pass, STAR was provided with genome indexes generated using the default settings. At the end of the first mapping pass, a new index was generated by combining splice junctions. Then, a second pass was performed using the new index generated by the first pass. The chimSegmentMin option was switched on for chimeric alignments.

2.2.2. Generation of Assembled Transcripts

The mapped reads obtained from the STAR aligner were used as input for StringTie to generate assembled transcripts for linear junction recognition. Annotations from known transcripts were used as a transcript model reference to guide the assembling process with the “-G” option. The StringTie options were as follows: `stringtie < input mapped_bam > -f 0.1 -o < file > -p 4 -G < GTF`.

![Assembling Splice Junctions Analysis](image)

**Fig. 1.** A schematic overview of the ASJA workflow. The ASJA architecture consists of three layers (from top to bottom) including chimeric alignment by STAR, junction identification by different model based on characteristics of the three types of junctions, and finally integration of splicing junctions utilization ratio and gene status.
2.3. Junction Extraction

2.3.1. Extraction of Linear Junctions

After mapping and assembling, a custom perl script was employed to extract linear splice junctions from the assembled transcripts as well as to calculate their expression levels using the formula, \( CPT(AB) = \min \left( \sum \text{cov}(A), \sum \text{cov}(B) \right) \times 10,000,000/TC \), where \( \text{cov}(A) \) and \( \text{cov}(B) \) represent the coverage of each end of the exon. We combined the coverages of all flanking exons for each junction and used the minimal expression level (5′ or 3′ flanking exon level). Junction expression was further normalized by the total annotated junction coverage (TC).

2.3.2. Extraction of Back-Splice Junctions

A custom perl script was used to extract back-splice junctions from chimeric alignments. Back-splice junction candidates require: 1. chimeric reads mapping to the same chromosome and strand; 2. the distance between the 3′ splice donor site and 5′ splice acceptor site to be <3,000,000 bp; 3. junction type = 1 (GT/AG) or junction type = 2 (CT/AC); 3. Back-splice junctions from the mitochondrial and other unannotated chromosomes, and those with unreasonable start and end positions were not considered.

2.3.3. Extraction of Fusion Junctions

Fusion junctions were also extracted from chimeric alignments. To reduce the false positive rate, the following steps were taken: 1. the junction reads mapping to mitochondrial and other unannotated chromosomes or unmapped contigs were filtered out; 2. chimeric reads with ‘junction type = 0’ (other motif expect GT/AG and CT/AC) were filtered out; 3. back-spliced reads (circRNAs) were excluded; 4. putative fusion junctions were extracted for further filtering when the supporting read number was >1. ASJA calculates the spanning reads as according to the parameter SpanningReads in STARChip [24]. Supporting and spanning reads can be helpful for checking the validity of output fusions.

2.4. Junction Annotation and Integration

2.4.1. Preparation of the Junction Annotation File

ASJA generates a primary annotation file for linear junctions as well as a filtered annotation file without read-through and paralog genes that have been removed from the reference gene annotation (GTF). To annotate back-splice and fusion junctions, ASJA provides an annotation of exons in BED (Browser Extensible Data) format for straightforward assessment of annotation status.

Fig. 2. Performance of ASJA on validation dataset. (A) Venn diagram shows the number of circRNAs predicted by three circRNA prediction tools, using 12 normal tissues. (B) Overlap of prediction results between two samples (RNaseR+, ribominus RNA treated with RNase R; RNaseR-, ribominus RNA) for three tools. (C) Times consumed (Minutes) by the softwares to analyse each run of validation samples.
Table 1
The performance of different fusion junction detectors using validated samples.

| Samples          | Statistics | ASJA  | MapSplice2 | deFuse |
|------------------|------------|-------|------------|--------|
| CGGA_661(1)*     | Total      | 3     | 24         | 35     |
|                  | TP         | 1     | 1          | 1      |
|                  | Recall     | 100%  | 100%       | 100%   |
|                  | Precision  | 33.30%| 4.16%      | 2.85%  |
| CGGA_374(3)      | Total      | 24    | 38         | 73     |
|                  | TP         | 2     | 2          | 3      |
|                  | Recall     | 66.70%| 66.70%     | 100%   |
|                  | Precision  | 8.30% | 5.20%      | 4.10%  |
| CGGA_1329(5)     | Total      | 48    | 71         | 153    |
|                  | TP         | 4     | 5          | 5      |
|                  | Recall     | 80%   | 100%       | 100%   |
|                  | Precision  | 8.33% | 7.04%      | 2.26%  |

Note: * The value in parentheses is the number of validated fusions.

2.4.2. Calculation of the Splice Ratio

For linear junctions, ASJA calculates the ratio of splice junctions by normalizing the maximum CPT of splice junctions within a gene as follows:

\[ \text{Weight ratio} = \frac{\text{CPT}}{\text{CPT}_\text{m}} \]

Where CPT is the maximum CPT of the gene the junction originated from, and CPTm is the maximum CPT of the same splice site as the donor.

ASJA calculates the ratio of splice junctions by the following formula:

\[ \text{donor ratio} = \frac{\text{donor read}}{\text{sum(donor read, linear read)}} \]

\[ \text{acceptor ratio} = \frac{\text{acceptor read}}{\text{sum(acceptor read, linear read)}} \]

Where linear read is the number of linear junction reads with the same splice site as the donor.

2.4.3. Integrative Analysis of the Three Types of Junctions

Based on gene annotation status and splice junction position, ASJA integrates the three types of junctions. The output context includes linear_junction_ID, gene_name, circRNA_ID and fusion_ID.

2.4.4. Junction Filtering

To get high confidence junctions, ASJA provides a script to filter junctions based on read count and ratio. We recommend the following settings: linear junctions with a reads count > 1 and a weight ratio > 0.08, back-splice junctions where the circRNA back-splicing read count is > 1, fusion junction with at least one fusion splice site to locate on the boundaries of known exons.

2.5. Datasets

2.5.1. Validation Datasets for ASJA Performance

The RNA-seq datasets from poly(A)- and poly(A)-/RNaseR RNAs in human PA1 cells were download from GEO (GSE75733) [30] for the comparison of back-splice junction detection. Three RNA-seq datasets from glioma (GBM) were downloaded from NCBI SRA, including SRR934794, SRR934744 and SRR934930 [31] for the comparison of fusion junction detection.

Table 2
The number of junctions in each sample.

| Sample          | Linear junction | Back-splice junction | Fusion junction |
|-----------------|-----------------|----------------------|----------------|
| Brain01         | 187,293         | 14,055               | 0              |
| Brain02         | 165,503         | 9542                 | 0              |
| Colon01         | 164,370         | 5365                 | 0              |
| Colon02         | 167,407         | 5736                 | 1              |
| stomach01       | 171,263         | 4226                 | 1              |
| stomach02       | 151,269         | 2733                 | 1              |
| Liver01         | 152,997         | 3920                 | 0              |
| Liver02         | 139,087         | 2894                 | 1              |
| Heart01         | 162,130         | 8084                 | 0              |
| Heart02         | 151,089         | 6351                 | 0              |
| Lung01          | 173,994         | 6304                 | 2              |
| Lung02          | 172,911         | 6169                 | 1              |
| BLCA_N          | 169,543         | 3465                 | 0              |
| BLCA_T          | 166,492         | 5936                 | 2              |
| BRCA_A          | 151,840         | 9875                 | 0              |
| BRCA_T          | 173,950         | 10,887               | 13             |
| CRC_N           | 172,163         | 5600                 | 3              |
| CRC_T           | 172,762         | 4696                 | 0              |
| GC_N            | 146,109         | 3469                 | 2              |
| GC_T            | 155,783         | 2278                 | 0              |
| HCC_N           | 163,502         | 5338                 | 0              |
| HCC_T           | 171,358         | 5513                 | 4              |
| KCA_N           | 167,979         | 6101                 | 1              |
| KCA_T           | 163,515         | 7895                 | 1              |
| PRAD_N          | 169,059         | 6070                 | 1              |
| PRAD_T          | 161,242         | 3955                 | 1              |
| Sum of unique junctions | 322,675         | 81,484               | 33             |

Note: BLCA: bladder urothelial carcinoma, BRCA: breast cancer, CRC: colorectal cancer, HCC: hepatocellular carcinoma, GC: gastric cancer, KCA: kidney clear cell carcinoma, PRAD: prostate adenocarcinoma.

2.5.2. Datasets for Application ASJA

RNA-seq datasets from twelve normal tissues, seven cancerous tissues and seven matched adjacent (NT) tissues from GEO (GSE77661) [28] were download to apply to ASJA. Normal tissues included brain, colon, heart, liver, lung and stomach, and the seven cancerous tissues contained bladder urothelial carcinoma (BLCA), breast cancer (BRCA), colorectal cancer (CRC), hepatocellular carcinoma (HCC), gastric cancer (GC), kidney clear cell carcinoma (KCA) and prostate adenocarcinoma (PRAD).

3. Results

3.1. Overview of ASJA

ASJA is a program to obtain, annotate, and integrate three types of splice junctions from reference-based assembled transcripts (linear junction) and chimeric alignments (back-splice and fusion junctions) (Fig. 1). ASJA works as follows: i) ASJA takes advantage of STAR aligner and StringTie assembler to generate chimeric alignments and assembled transcripts from RNA-seq reads respectively. ii) ASJA detects and quantifies linear junctions based on the assembled transcripts. Back-splice and fusion junctions were extracted from the chimeric alignments. iii) ASJA provides quantification and annotation of the three junction types and produces an integrated file accounting for their relationships.

3.2. Performance Evaluation

3.2.1. Linear Junction

The gold standard of known splice junction candidates were defined by the annotated junctions (known genes) from 1-pass alignment with expression read count > 1 for junction and FPKM > 0.1 for corresponding gene (colon01, total: 20,618) [32]. The sensitivity of ASJA known linear junctions is 97.3%. For novel linear junctions, the sensitivity is 89.8% from result of comparing the known splice of 2-pass without annotation.
with gold standard. Moreover, we downloaded MapSplice2 software which designs for mapping RNA-seq read to reference genome for splice junction discovery [33]. The sensitivity is 91.5% when we evaluated known splice junctions of Mapsplice2 against with gold standard.

3.2.2. Back-Splice Junction

For back-splice junction (circRNA) detection, we performed a comparison of ASJA with two other algorithms, circRNA_finder [12] and ACFS [25]. We used the RNA-seq datasets from 12 normal tissues (GSE77661). Of the three tools, ASJA detected the highest number of circRNAs. The high proportion of identical circRNAs (75.5%) was observed among the three tools (Fig. 2A). To assess the false positive rate, we used the RNase R (for validation of circRNA) digested RNA-seq and corresponding poly(A)-RNA-seq datasets (GSE75733) to explore the level of false positive circRNAs. We observed ASJA also has a similar false positive rate (31.2%) compared to the other algorithms, including ACFS (43%) and circRNA_finder (31.5%) (Fig. 2B).

3.2.3. Fusion Junction

For fusion junction detection, we made the comparison with two other fusion detectors, including MapSplice2 and deFuse [34]. We used three validation samples from GBM as positive sets, including SRR934794, SRR934744 and SRR934930. A total of 9 fusion genes has been validated in the three RNA-seq datasets [31]. We observed that ASJA has more precision rate than the other softwares, although they all have high recall (Table 1).

3.2.4. Running Time

We also compared the running time of ASJA with other tools. ASJA shows an excellent performance in computational time, which is 2–10

---

**Fig. 3.** The characteristics of ASJA identified linear junctions in human cells. (A) The pie chart shows the number of raw (grey) and high confidence (red) linear junctions. (B) The distribution of the linear junctions to annotated known genes. The unannotated junctions are shown as novel junctions. (C) The pie chart shows the proportion of gene isoforms and intergenic junctions in novel junctions. (D) Bar chart shows the number of novel junctions in different normal tissues. Gene isoforms are shown in red. Intergenic genes are shown in blue. (E) Clustering analysis of all differentially expressed junctions between cancer tissues and NCTs. The heatmap is based on expression values with log2(fold-change) > 1 and p < .01 (Wilcoxon test).
fold faster than others (Fig. 2C). The time starts from the processing of the FASTQ file until the generation of final candidates, including linear, back-splice and fusion junctions.

3.3. Application to Various Samples

3.3.1. Detection of Splice Junctions in RNA-Seq Data from Normal and Cancerous Tissues by ASJA

We implemented ASJA to detect and quantify the splice junctions in our recent published RNA-seq data (GSE77661), which contains 12 normal tissues, 7 cancerous tissues and 7 matched adjacent (NT) tissues. In total, we detected 322,675 linear junctions, 81,484 back-splice junctions and 33 fusion junctions after removing duplicate junctions detected in the data set (Table 2). Each sample contains a median of 165,997 linear junctions, 5668 back-splice junctions, and 1 fusion junction. The number of linear junctions is significantly higher than the number of back-splice and fusion junctions. The comparison of read counts between back-splice and linear junctions showed that the ratio of back-splice junctions was about 1% except in brain tissue (4%). This result is consistent with the notion that circular RNAs are generally enriched in tissues where mitotic division is not prevalent [12]. In addition, fusion junctions were hardly detected in normal tissues while there were some fusion junctions in cancers such as BRCA.

3.3.2. The Characteristics of ASJA Detected Linear Junctions in Human Cells

Linear junction calls were performed using the default ASJA settings. In total, 322,675 distinct linear junctions were found in all tissues with 284,287 of these junctions containing at least two unique read counts and a weight ratio >0.08 (Fig. 3A). These linear junctions were derived from 100,774 known primary transcripts and categorized into four types according to their genomic origin. >78.5% of the linear junctions were located in protein-coding regions, whereas smaller fractions aligned with long noncoding RNAs, and pseudogenes with known transcripts (Fig. 3B). Moreover, we observed an average of 10.4 known linear splicing junctions per gene, and noted that 10,870 of 240,453 known linear junctions are in the 5’ noncoding regions of mRNAs. Of note, 43,834 (15.6%) of the linear junctions are unannotated. Most of the novel junctions (33,480) overlapped with known genes, while the remaining 10,353 (23.6%) junctions originated in intergenic regions (Fig. 3C). Our results revealed that numerous novel junctions seem to...

Fig. 4. The characteristics of ASJA identified back-splice junctions (circRNA) by in human cells. (A) The pie chart shows the number of annotated and unannotated circRNAs. (B) Bar plots showing the number of genes in different number of back splicing events. (C) Genomic origin of circRNAs. The pie chart shows the genomic distribution of all predicted/annotated circRNAs. (D) Multidimensional scaling screen for the identification highly abundant circRNAs. Red dots and grey dots represent highly abundant and low-abundance circRNAs, respectively.
be specifically expressed in various tissues. There were 7107 novel linear junctions in brain tissue, which is much higher than in other tissues (Fig. 3D). We also compared the junction expression profiles of cancer tissues and matched non-cancer tissues (NCTs), and identified 109 downregulated and 765 upregulated junctions in cancer (Fig. 3E).

3.3.3. The Characteristics of ASJA Detected back-Splice Junctions (circRNA) in Human Cells

Of the 31,346 circRNAs identified in all samples (26 tissues) by ASJA, 20,475 were not yet annotated in circBase (Fig. 4A). There are only 2 median, ranging from 1 to 72) back-splicing event per gene (Fig. 4B). We investigated the genomic origin of these circRNA candidates using Genencode references. >90% of the circRNAs consisted of protein-coding exons, while smaller fractions aligned to long noncoding RNAs, and antisense regions of known transcripts (Fig. 4C). ASJA quantified the abundance of each circRNA with respect to its alternative linear isoform by estimating the back-splice ratio at the 5’end or 3’ end. Although linearly splice products were absent in some cases, the back-splice ratio for these sites varied considerably. When using a stringent back-splice ratio and read count cut-off (mean of back ratio > 0.15; log2 (average of circRNA supporting read) > −1), we observed 404 high-abundance circRNAs (Fig. 4D).

4. Discussion

High throughput RNA-seq technologies have given rise to large amounts of data that bring about an unprecedented challenge for the development of computational tools. ASJA is developed to detect and quantify different splice junctions from RNA-seq data. ASJA is useful not only for cataloging linear, back-splice and fusion junctions in samples but also for the identification of novel junctions. ASJA can quickly and accurately identify the different kinds of junctions and evaluates their expression levels, allowing for direct comparisons across different samples.

We compared the accuracy and speed of ASJA to several other splice junctions detecting programs, including MapSplice2, circRNA_finder, ACPS, using validation datasets. We showed that ASJA has better precision, especially for fusion genes. Most fusion gene detection softwares have higher recall compared to ASJA, however the false positives are very high, which requires further complex screening to obtain the real fusion genes. Moreover, the speed of ASJA is much faster than that of others. In brief, ASJA can get more accurate junction information in a shorter time.

Through comprehensive real data analysis, we prove that at least three major improvements of our software: (i) compared with existing common methods for junction identification based on RNA-seq data, the proposed ASJA procedure can simultaneously identify three kinds of junction for large sample sizes, while previous software could only identify a maximum of two kinds of junction. (ii) ASJA extracted linear splice junction from reference-based assembled transcripts to control false-positive rate (FDR) at desired levels, while competing methods have exceedingly inflated FDR levels when obtain linear junction based on mapped reads. (iii) our method can integrate multiple information about a sample including splicing junctions, gene annotation, read count, normalized expression level and splice ratio of each junctions for downstream analysis.

The ASJA pipeline and its default parameters are designed for simultaneously detecting three kinds of junctions. However, this procedure may miss splicesome as result of failing to consider some special circumstances. ASJA will miss a junction since it concerned GT-AG dinucleotides regulation, so junctions with non-canonical donor and acceptor sites will be ignored. Although ASJA can discover novel transcripts based on junction signal, it is limited to pre-RNA containing multiple exons and is not applicable to single-exon. Further investigation is emphasized to understand how the filter should be adjusted to reduce false positive and how to account for other factors like sequencing depth and coverage ratio or variability in junction profiles between samples.

5. Conclusions

In conclusion, ASJA is a powerful tool for the detection and characterization of different kinds of splice junctions including novel junctions. This method greatly facilitates the identification and cataloguing of splice junctions from RNA-seq data and offers new possibilities for exploring transcriptome complexity. As an application of this method, we have made available package for the research community to use and will regularly update it.

Authors’ Contributions

SH, JZ and XH proposed the initial idea and designed the methodology, JZ, QL and YL implemented the concept and processed the results. SH, JZ, QL and OZ wrote the manuscript. All authors read and approved the final manuscript.

Declarations of Competing Interest

None.

Availability

The current version of ASJA can be downloaded from https://github.com/HuangLab-Fudan/ASJA

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (81702786, 8187229, 81672779) and the Shanghai Sailing Program (17YF140260).

References

[1] Montes M, Sanford BL, Comiskey DF, Chandler DS. RNA splicing and disease: animal models to therapies. Trends Genet 2019;35:68–87.
[2] Scotti MM, Svanson MS. RNA mis-splicing in disease. Nat Rev Genet 2016;17:19–32.
[3] Mertens F, Johansson B, Florestos T, Mittelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15:371–81.
[4] Mittelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat Genet 2004;36:331–4.
[5] Mittelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007;7:233–45.
[6] Westbrook CA, Hoehner AL, Spino C, Dodge RK, Larson RA, Davey F, et al. Clinical significance of the BCR-ABL fusion gene in adult acute lymphoblastic leukemia: a Cancer and leukemia group B study (8762). Blood 1992;80:2983–90.
[7] Druker BJ, Guilhot F, O’Brien SG, Gathmann I, Kantarjian H, Gattermann N, et al. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. N Engl J Med 2006;355:2408–17.
[8] Tkachuk DC, Westbrook CA, Andreuff M, Donlon TA, Cleary ML, Suryanarayan K, et al. Detection of bcr-abl fusion in chronic myelogenous leukemia by in situ hybridization. Science 1990;250:559–62.
[9] Solomon B, Varella-Garcia M, Camidge DR. ALK gene rearrangements: a new therapeutic target in a molecularly defined subset of non-small cell lung cancer. J Thorac Oncol 2009;4:1450–9.
[10] Mehra R, Tomlins SA, Shen R, Nadeem O, Wang L, Wei JT, et al. Comprehensive assessment of TPMRSS2 and ETS family gene aberrations in clinically localized prostate cancer. Mod Pathol 2007;20:538–44.
[11] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehta R, Sun XW, et al. Recurrent fusion of TPMRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8.
[12] Menczak S, Jens M, Elsfiohni A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potential. Nature 2013;495:333–8.
[13] Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. Nature 2013;495:384–8.
[14] Wilusz JE, Sharp PA. Molecular biology. A circuitous route to noncoding RNA. Science 2014;340:440–1.
[15] 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;56:55–66.
[16] Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, et al. Exon-intron circular RNAs regulate transcription in the nucleus. Nat Struct Mol Biol 2015;22:256–64.
[17] Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, et al. Circ-ZNF609 is a circular RNA that can be translated and functions in Myogenesis. Mol Cell 2017;66:22–37.e9.
[18] Yang Y, Gao X, Zhang M, Yan S, Sun C, Xiao F, et al. Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. J Natl Cancer Inst 2018;110.
[19] Li X, Yang L, Chen LL. The biogenesis, functions, and challenges of circular RNAs. Mol Cell 2018;71:428–42.
[20] Engstrom PG, Steiger T, Siposs B, Grant GR, Kahles A, Ratsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 2013;10:1185–91.
[21] Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernandez J, Collado-Torres L, Wang S, et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the sequence read archive. Genome Biol 2016;17:266.
[22] Hansen TB, Veno MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res 2016;44:e58.
[23] Chen B, Huang S. Circular RNA: an emerging non-coding RNA as a regulator and biomarker in cancer. Cancer Lett 2018;418:41–50.
[24] Akers NK, Schadt EE, Losic B. STAR chimeric post for rapid detection of circular RNA and fusion transcripts. BIOINFORMATICS 2018;34:2364–70.
[25] You X, Conrad TO. Arcs: accurate circRNA identification and quantification from RNA-Seq data. Sci Rep 2016;6:38820.
[26] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. BIOINFORMATICS 2013;29:15–21.
[27] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33:290–5.
[28] Zheng Q, Bao C, Guo W, Li S, Chen J, Chen B, et al. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. Nat Commun 2016;7:11215.
[29] Li S, Hu Z, Zhao Y, Huang S, He X. Transcriptome-wide analysis reveals the landscape of aberrant alternative splicing events in liver cancer. HEPATOLOGY 2019;69:359–75.
[30] Zhang XQ, Dong R, Zhang Y, Zhang J, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res 2016;26:1277–87.
[31] Bao ZS, Chen HM, Yang MY, Zhang CB, Yu K, Ye WL, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 2014;24:1765–73.
[32] Veeneman BA, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. Two-pass alignment improves novel splice junction quantification. Bioinformatics 2016;32:43–9.
[33] Wang X, 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:e178.
[34] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 2011;7:e1001138.