BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 02/06 pp15–31
Volume 5, Number 1, March 2022
DOI: 10.26599/BDMA.2021.9020014

A Comparison of Computational Approaches for Intron Retention Detection

Jiantao Zheng†, Cuixiang Lin†, Zhenpeng Wu, and Hong-Dong Li*

Abstract: Intron Retention (IR) is an alternative splicing mode through which introns are retained in mature RNAs rather than being spliced in most cases. IR has been gaining increasing attention in recent years because of its recognized association with gene expression regulation and complex diseases. Continuous efforts have been dedicated to the development of IR detection methods. These methods differ in their metrics to quantify retention propensity, performance to detect IR events, functional enrichment of detected IRs, and computational speed. A systematic experimental comparison would be valuable to the selection and use of existing methods. In this work, we conduct an experimental comparison of existing IR detection methods. Considering the unavailability of a gold standard dataset of intron retention, we compare the IR detection performance on simulation datasets. Then, we compare the IR detection results with real RNA-Seq data. We also describe the use of differential analysis methods to identify disease-associated IRs and compare differential IRs along with their Gene Ontology enrichment, which is illustrated on an Alzheimer's disease RNA-Seq dataset. We discuss key principles and features of existing approaches and outline their differences. This systematic analysis provides helpful guidance for interrogating transcriptomic data from the point of view of IR.

Key words: alternative splicing; intron retention; gene expression; RNA-Seq

1 Introduction

Alternative Splicing (AS) is a common phenomenon in eukaryotes[1]. For genes with multiple exons in the human genome, 95% are alternatively spliced[2]. AS is an important transcriptional regulation mechanism that increases the structural and functional diversity of gene products[3,4]. Several studies have found that AS is associated with cancers and other complex diseases[5–8]. AS can be divided into five modes: exon skipping, alternative 3' splice site, alternative 5' splice site, mutually exclusive exons, and Intron Retention (IR). IR has received the least attention because Intron-Retained Isoforms (IRIs) were previously thought to be the consequence of mis-splicing of pre-mRNAs[9].

Previous studies have explored IR from the perspective of its location and conditions of occurrence. Galante et al.[10] conducted a large-scale IR analysis of 21 106 known human genes. They found that 14.8% of genes have at least one IRI and most IRs are located within UnTranslated Regions (UTRs). Considering species differences, a comparative analysis of humans and mice indicates that at least 22% of IRs in humans also exist in mice. In 2007, Sakabe and De Souza[11] proposed that the occurrence of high-frequency IR in humans is generally accompanied by the following conditions: (1) genes with short intron lengths; (2) high expression levels; (3) weak splice sites strength; (4) low density of exonic splicing silencers; and (5) low

• Jiantao Zheng, Cuixiang Lin, Zhenpeng Wu, and Hong-Dong Li are with the Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: jiantao@csu.edu.cn; lincxcsu@csu.edu.cn; zhenpeng@csu.edu.cn; hongdong@csu.edu.cn.

† Jiantao Zheng and Cuixiang Lin contributed equally to this paper.

* To whom correspondence should be addressed.

Manuscript received: 2021-04-22; revised: 2021-08-09; accepted: 2021-08-20

density of the intronic splicing enhancer GGG. Then Louro et al.[12] explored whether or not long introns are expression noise or expression choice in noncoding RNA transcription. In 2010, Cenik et al.[13] performed an in-depth analysis of human 5' UTR introns. Their study have shown that 5' UTR introns, which are generally in highly expressed genes, are relatively short. These introns are not completely absent from lowly expressed genes but rather enhance their expression in a length-dependent manner[14].

In recent years, some studies have discovered the regulatory role of IR in gene expression and its association with complex diseases. For example, Zhang et al.[15] explored the effect of IR in lung adenocarcinoma. They found that genes with tumor-specific IR are likely to be overexpressed in tumors and are likely to be lung cancer driver genes. IR may suppress the overexpression of genes that promote cancer development by triggering the Nonsense-Mediated Decay (NMD) mechanism[16, 17]. Using the RNA-Seq and exome data of 1812 cancer patients, Jung et al.[18] found that somatic Single-Nucleotide Variants (SNVs) causing IR are enriched in tumor suppressors. Most of these SNVs produce premature termination codon, resulting in loss of function by triggering NMD degradation or producing truncated proteins. These findings suggest that IR is a common mechanism of tumor suppressor inactivation. Furthermore, Dvinge and Bradley[19] explored the contribution of IR to the transcriptional diversity of many cancers. They found that the most common abnormal splicing mode in cancer is IR and that almost all types of cancers show frequent IR in cancer tissues.

As a mature sequencing technology, RNA-Seq has been widely used in the genome-wide analysis of pre-mRNA AS[20–22]. Powerful algorithms for IR detection have been developed to date. For example, Bai et al.[23] proposed two methods to detect IR events; one is IRcall based on ranking strategy, and the other is IRclassifier based on random forest classifier. These methods mainly consider gene expression information, read coverage within an intron, and read counts. They showed that their methods could effectively filter out false positives and improve the prediction accuracy of IR events in the Arabidopsis thaliana control experiment. The strong point of IRcall and IRclassifier is that they build a machine learning framework that can improve IR detection by introducing mature classification algorithms and exploring other features. Pimentel et al.[24] developed Keep Me Around (KMA) to find IR events. This method combines biological replicates to reduce the possibility of false-positive IR events and is well compatible with existing RNA-Seq quantification pipelines, such as eXpress[25]. The advantage of KMA is that it provides a way to realize the quantification of IR isoforms. In 2017, Middleton et al.[26] proposed IRFinder, which provides a complete end-to-end pipeline for the IR analysis of mRNA sequencing data. IRFinder uses read counts of introns to estimate the read coverage difference between the experimental and control groups through the Audic and Claverie Test. The main steps include sequence alignment via the STAR algorithm, quality controls on the sample analyzed, IR detection, and quantitative and statistical comparison of IR ratios between multiple samples. The end-to-end analysis designed by IRFinder is a good point. In our previous work, intron Retention Analysis and Detector (iREAD) was proposed[27] to detect IRs from RNA-Seq data. The main step is to screen out independent introns from the transcriptome annotation that do not overlap any exons of any isoforms. Then, it takes BAM files as input and calculates intron metrics, such as read counts, junction reads, Fragments Per Kilobase of exon model per Million mapped fragments (FPKM), and normalized entropy score. iREAD also allows users to manually add known IRs to the independent intron list for analysis. The principle of iREAD is simple and direct, and the concise use of commands is also one of its advantages.

In addition to the methods mentioned above for IR detection, the methods for detecting AS events can also be used for IR detection. For example, Mixture-of-ISOforms (MISO) proposed by Katz et al.[28] quantifies the expression level of AS genes from RNA-Seq data at the event level, including IR or isoform level, and provides confidence intervals of level estimation. MISO can also detect differentially regulated isoforms or exons between samples. The strength of MISO is that it can quantify AS events at the IR and isoform levels and provide confidence scores. Multivariate Analysis of Transcript Splicing (MATS)[29] proposed by Shen et al. analyzes differential AS patterns on non-replicated RNA-Seq data based on the Bayesian statistical framework. MATS works with almost any type of null hypothesis and supports user-defined thresholds to detect differential AS events. Replicate MATS (rMATS)[30], as improved MATS, uses a more robust hierarchical model to detect differential AS events from replicate RNA-Seq data. rMATS considers not only the sampling uncertainty in individual replicates, but also the variability among replicates. The advantage of rMATS is that it supports flexible hypothesis testing methods to detect differential AS events. DEXSeq[31] employs a Generalized Linear Model (GLM) to detect genes affected by Differential Exon Usage (DEU) with high sensitivity from RNA-Seq data. DEXSeq considers biological variation to achieve the reliable control of false discoveries. By replacing exon coordinates with intron coordinates, DEXSeq can also be used to detect IR[32]. However, this conversion method is not rigorous, which may result in the loss of some IR-related information. Unlike the method described above, Comprehensive Alternative Splicing Hunting (CASH)[33] detects AS events neither from incomplete reference genomes nor from transcripts built by third-party programs. CASH consists of two main phases: SpliceCons (splice site construction) and SpliceDiff (differential AS detection). By fully reconstructing the AS site from RNA-Seq data, SpliceCons significantly enhances the ability of CASH to detect novel AS events. To reduce false positives, SpliceDiff combines two statistical tests, including the test of AS inclusion/exclusion event expression and the test of AS exons relative to the entire gene expression. SpliceDiff also significantly improves the ability of CASH to detect AS events differentially between RNA-Seq samples. The advantage of CASH is that it does not rely on transcriptome annotation, which facilitates the discovery of novel IR events.

These methods differ in their features and performance in detecting IR. The principles of these methods have been systematically reviewed[34]. However, an experimental comparison of these methods is currently lacking. In this work, we perform a systematic experimental comparison of rMATS, DEXSeq, CASH, MISO, IRFinder, and iREAD. These methods are selected because they represent the state-of-the-art and widely used methods in their respective categories. IRFinder and iREAD are the two state-of-the-art methods dedicated to IR detection, whereas rMATS, DEXSeq, CASH, and MISO are the most widely used methods for the analysis of AS events, including IR. Considering that a gold standard dataset of IR is available, we use simulated data with known retained and nonretained introns to evaluate the predictive performance of each method using the Area Under the receiver operating characteristic Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC). These methods are currently only applicable to next-generation sequencing data and cannot directly support third-generation sequencing data. Thus, their results were compared with real next-generation sequencing data. Using an Alzheimer's disease RNA-Seq dataset as an example, we performed differential expression for the IRs obtained by each method and compared the Gene Ontology (GO) enrichment results. Finally, we discussed the advantages and limitations of each method and concluded with potential ways to improve IR detection.

2 Material and Method

2.1 Description of IR detection methods

As the field of IR is receiving increasing attention, many methods can be used to detect IR. The six methods included in this study are MISO, DEXSeq, rMATS, CASH, iREAD, and IRFinder. A brief description of each method is shown in Table 1.

(1) MISO

MISO (v0.5.4)[28] treats the expression levels of all isoforms in a gene as random variables whose distribution could be estimated using Markov Chain Monte Carlo (MCMC). MCMC sampler uses a Metropolis-Hastings sampling scheme, combined with a Gibbs sampling step. The calculation of isoform expression level estimation between genes is independent. Thus, MISO provides the basic functions

Table 1 Summary description of IR detection methods.

Method	Input	Output	Feature	Reference	Source code	Version
MISO	SAM/BAM	Percent spliced in	MCMC sampler; high degree of parallelization	[28]	Python, C	v0.5.4
DEXSeq	SAM/BAM	Number of reads per intron	Exon counting bins; DEU	[31]	Python, R	v1.24.4
rMATS	FASTQ/BAM	Inclusion level	Hierarchical model; likelihood-ratio test	[30]	Python, C, Cython	v4.0.2
CASH	BAM	Percent spliced in	User-friendly visual interface; SpliceCons and SpliceDiff	[33]	Java	v2.2.0
iREAD	BAM	FPKM, entropy, junction read counts, all read counts	Independent intron; normalized entropy	[27]	Python, Perl	v0.8.0
IRFinder	Fastq/BAM	IR-ratio	Complete pipeline for IR analysis of multi-species mRNA sequencing data	[26]	C, Perl	v1.2.6

of running on a cluster to achieve a high degree of parallelization.

The differential analysis module of MISO relies on the single sample analysis results of the previous stage. Thus, MISO cannot directly handle experiments with multiple biological replicates. One solution is to merge all BAM files in groups and then input them into the MISO pipeline. Another solution is to extract the assigned reads of isoforms in the single sample analysis stage as input to other differential analysis software, such as empirical analysis of digital gene expression data in R (edgeR) and Differential Expression analysis for Sequence count data (DESeq2). MISO accepts a genomic annotation file from any source (e.g., RefSeq, UCSC, or Ensembl) as long as it is organized in the General Feature Format (GFF3). As for paired-end data, we need to provide additional information, such as mean insert length and standard deviation which can be obtained through the pe_utils module of MISO.

(2) DEXSeq

DEXSeq (v1.24.4)[31] uses a GLM to detect genes affected by DEU with high sensitivity from RNA-Seq data. The core data structure of DEXSeq is a table that records exon-overlapping reads. The table is obtained mainly by the following steps. First, DEXSeq defines a set of coordinate interval lists of exons, called "exon counting bins" to divide exons. When exons appear on different boundaries of different transcripts, exon counting bins may also indicate a part of exons. Second, for each BAM file, DEXSeq counts the number of reads that overlap with exon counting bins to obtain the table above. Third, DEXSeq fits a GLM for each gene to test DEU. DEXSeq provides a function interface for differential analysis, which encapsulates DESeq2. Users can organize the data into the required format to perform differential analysis. In particular, DEXSeq considers biological variation to achieve the reliable control of false discoveries. By replacing exon coordinates with intron coordinates, DEXSeq can also be used to detect IR[32].

(3) rMATS

rMATS (v4.0.2)[30] constructs a hierarchical model to estimate the exon inclusion level (ϕ) using reads that uniquely map to exon inclusion or skipping isoforms. However, the estimation of ϕ is affected by many factors. For example, in a single sample, the estimation of ϕ is affected by the sequencing coverage of AS events, and higher sequencing coverage leads to more reliable estimates. In the sample group, due to biological or technical reasons, the differences between replicates also affect the estimation of ϕ. Correspondingly, given that different isoforms have different lengths, rMATS normalizes the reads mapped to the exon inclusion or skipping isoform by the effective length of the isoform when calculating ϕ, which solves the first problem mentioned above. Differences between replicates could be resolved through estimating the group mean of ϕ as fixed effects followed by conducting mixed modeling of replicates within the group. Compared with the classic hypothesis that compares the equality of ϕ between sample groups, rMATS uses a likelihood ratio test to analyze whether or not the difference in the mean ϕ between groups exceeds a user-defined threshold.

(4) CASH

Many known methods (such as rMATS) rely on annotated transcripts, but current transcript annotations are incomplete. Some methods intend to assemble transcripts from third-party programs, such as Cufflinks, but the accumulation of false transcripts during the assembly may lead to the false prediction of AS events. In addition, even some transcripts containing novel AS events may have an excessively low expression to construct. The current third-generation sequencing can solve the above problems, but its high cost restricts its wide application in transcriptome studies. Considering the existing problems mentioned above, CASH (v2.2.0)[33] proposed two key components, SpliceCons and SpliceDiff, which significantly improve the predictive ability of AS events.

SpliceCons extracts all non-redundant junction reads from the RNA-Seq data and combines the annotated exon sites from the reference genes to construct all the splice sites for each gene to detect the novel AS event. SpliceDiff computes the False Discovery Rate (FDR) using the Benjamini Hochberg method to reduce false positives.

The input of CASH includes BAM files that are required to be sorted and indexed and annotation files in gene transfer format or GFF3. The output of CASH contains not only the Percent Spliced In (PSI) value of introns/exons but also intergroup information, such as delta-PSI, P-value, and FDR.

(5) iREAD

iREAD (v0.8.0)[27] assumes that reads for retained introns are evenly distributed throughout the intronic region. Thus, iREAD calculates the entropy score to detect IR events that do not overlap with exons from poly-A enriched RNA-Seq data. The statistical information of intron mainly includes the Total number of Reads (TR), the number of exon-intron Junction Reads (JR), the FPKM value, and the normalized entropy score (NE-score). iREAD defines a set of high threshold values (TR⩾20, JR⩾1, FPKM⩾3, and NE⩾0.9) to detect IR events strictly.

iREAD accepts a sequence alignment file in BAM format and an annotation file of independent introns in BED format as input. iREAD also needs to input the number of reads mapped to the whole genome to calculate FPKM, which can be obtained through SAMtools (v1.6)[35]. Although iREAD does not provide the functionality of differential analysis, the read counts in its output file can be directly used as the input of differential analysis software, such as edgeR or DESeq2.

(6) IRFinder

IRFinder (v1.2.6)[26] provides a complete pipeline for the IR analysis of mRNA sequencing data. IRFinder uses IR-ratio to measure IR in terms of splicing level, which reflects the proportion of intron retaining transcripts. To reduce the impact of noise on IR estimates, IRFinder excludes regions where each intron is covered by highly expressed features, such as snoRNAs, microRNAs, or unannotated exons. IRFinder also filters out samples that are not suitable for IR detection, such as samples that suffer from high levels of DNA contamination or are mislabeled as mRNA sequencing. The differential analysis module of IRFinder captures the IR-ratio change of introns across multiple samples and calculates the significance. Considering the difference in the number of biological replicates, IRFinder provides different ways to perform differential analysis. If the number of replicates between groups is less than four, the Audic and Claverie Test should be used; otherwise, Student's t-test can be used. IRFinder also provides modules that encapsulate DESeq2 for differential analysis.

IRFinder has some special requirements for input files. For example, the transcriptome annotation file and genomic sequence file require fixed names of "transcript.gtf" and "genome.fa", respectively. The input BAM file must be sorted according to the read name.

2.2 RNA-Seq datasets

We use one simulation dataset and two real RNA-Seq datasets. The description on each dataset is shown in Table 2. Details are as follows.

2.2.1 Simulation data

Considering the unavailability of a gold standard dataset of IR, we used simulated data with known retained and nonretained introns to evaluate the predictive performance of each method using AUC and AUPRC. The simulation process is described below. First, we used the Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEER)[36] to generate paired-ended RNA-Seq simulation data of the mouse genome (version: mm10). Following the advice of the author of BEER, we turned off the option for generating novel splice isoforms to facilitate IR analysis. We first simulated a dataset of high sequencing depth of 60 million reads with a read length of 100 bp, which is called SIMU60. We then generated another two datasets with 15 and 30 million reads by sampling from SIMU60, which are called SIMU15 and SIMU30 for short.

Table 2 Description for three mouse RNA-Seq datasets.

Dataset	Read length (bp)	Sequencing depth ($\times 10^5$)	Source
Simulation data	100	15, 30, 60	BEER
Upf2 knockout mouse data	75	20	GEO (accession ID: GSE26561)
App mutant mouse data	101	100	Tau and APP mouse model study

In the subsequent comparative analysis, we focused on investigating the IR prediction performance of each method on SIMU30 with medium sequencing depth. The gene numbers of SIMU15, SIMU30, and SIMU60 were 22 202, 22 298, and 22 376, respectively, and their intron numbers were 64 161, 69 338, 73 544, respectively. These simulated data were mapped to the Ensembl mouse transcriptome annotation (version 75) and genome (GRCm38) using STAR (v2.6.0a) with default settings to generate aligned reads in BAM files.

2.2.2 RNA-Seq data

The first dataset is from a control experiment of liver tissue from the Upf2 knockout mouse as described in IRFinder without biological replicates (accession ID: GSE26561)[26]. It contains two single-end raw sequencing data in SRA format, with a read length of 75 bp and a sequencing depth of 20 million reads. The dataset was divided into a control group (Wild-Type, WT) and an experimental group (KnOckout, KO). These two SRA files were first converted to FASTQ files by using fastq-dump (v2.9.2). Then, the reads were mapped to the Ensembl mouse transcriptome annotation (version 75) and genome (GRCm38) using STAR (v2.6.0a) with default settings, and BAM files with aligned reads were obtained.

The second dataset was generated from APPPS1 mouse models of amyloidosis, which was produced in the Accelerating Medicines Partnership-Alzheimer's Disease Project (https://www.synapse.org/#!Synapse:syn17008852). Brain samples were sequenced using Illumina HiSeq 2000 to obtain paired-end sequence data with a sequencing depth of approximately 100 million and a read length of 101 bp. We selected 16 female samples from the APP mice in the study. Among them, eight normal samples served as the control group, and eight APP mutant samples served as the experimental group.

3 Result

Intron annotations adopted by each method are different. Hence, the number of shared introns among more than three methods is small. For example, iREAD custom independent intron and CASH reconstruct AS site. We addressed this problem by using "intron cluster", which is defined as non-overlapping intron regions of all genes as proposed in LeafCutter[37]. In this way, we can compare these methods at the cluster level by mapping the detected retained introns of each method to the intron cluster to obtain the intersection. The specific steps for generating intron clusters are as follows. First, GTFtools[38] was used to obtain genome-wide introns in BED format from the genome annotation file. Then, BEDtools (v2.25.0)[39] were employed to merge the overlapping introns by gene, and an intron cluster was obtained. With regard to the rules for mapping to intron clusters, the target intron with the largest overlap with a given intron was selected.

3.1 Performance comparison on simulated data

In this section, we compared the IR prediction performance of each method on SIMU30 data. rMATS and CASH analyze AS events differentially and require two groups of samples. Therefore, we generated SIMU50 with 50 million reads sampled from SIMU60 to form a group with SIMU30 to meet the input requirements of rMATS and CASH. Based on the configuration, BEER was used to generate the gold standard dataset of IR. For each intron, the TR, FPKM, and the number of JR[36] were calculated. The threshold values (TR$\geqslant$10, FPKM$\geqslant$0.3, and JR$\geqslant$1) were used to define IR events following the work in Ref. [27]. SIMU30 has a total of 69 338 introns, of which 20 332 were retained. Then, we labeled the intron cluster based on the labels of the corresponding simulation data. For example, for intron A in the intron cluster, we selected intron B from the simulation data that has the most overlaps with A's coordinates and assigned the label of B to A.

Then, we run each method to obtain their predicted IR events. The criteria to define IR for each method are described in the following. MISO, rMATS, IRFinder, and CASH were used to estimate a metric related to IR levels, namely, PSI, inclusion level, IR-ratio, and PSI. We set the threshold value of 0.1 to filter for IR events. iREAD directly labels whether or not each intron is retained based on the default setting. DEXSeq counts the number of reads that overlap with exon counting bins in SIMU30. We calculated FPKM based on the read count and set the same threshold value as used by iREAD to determine IR events.

Basing on the intron cluster, we extracted the intersection of IRs detected by all methods and used it to calculate AUC and AUPRC. IRFinder, MISO, DEXSeq, iREAD, CASH, and rMATS predicted 13 935, 12 298, 3237, 2650, 2114, and 630 IR events, respectively. IRFinder and MISO predicted tens of thousands of IRs, whereas rMATS predicted less than 1000 IRs. The difference in the number of IR is related to the detection criteria of each method. One possible reason is that IRFinder detects introns that appear in more than 10% of transcripts (IR-ratio$\geqslant$0.1). MISO calculates PSI at the gene level. rMATS may be related to the insufficient number of biological replicates, which leads to insufficient differential analysis between groups. The intersection of IR sets on SIMU30 is shown in Fig. 1, and other simulation data are shown in Figs. A1 and A2 in the Appendix. The prediction results of IRFinder and iREAD, two methods specially designed for IR, were different. Combining IR-ratio and FPKM, iREAD and IRFinder had different detection tendencies for IR. iREAD prefers IR with relatively higher expression, whereas IRFinder prefers IR with relatively higher IR-ratio[27]. The reason may be that iREAD only recognizes IR that is not found in the annotation file by default, and the default threshold for the NE of iREAD is high (NE$\geqslant$0.9). Similarly, DEXSeq, which detects IR based only on FPKM, predicted more than iREAD. Furthermore, the prediction results of rMATS and CASH were significantly different. The reason may be that CASH considers the annotation splicing site of the reference gene and the new splicing site of the RNA-Seq alignment data, whereas rMATS relies more on annotated transcripts of reference genes.

Given the algorithmic difference between the six methods, the number of shared IR events detected by

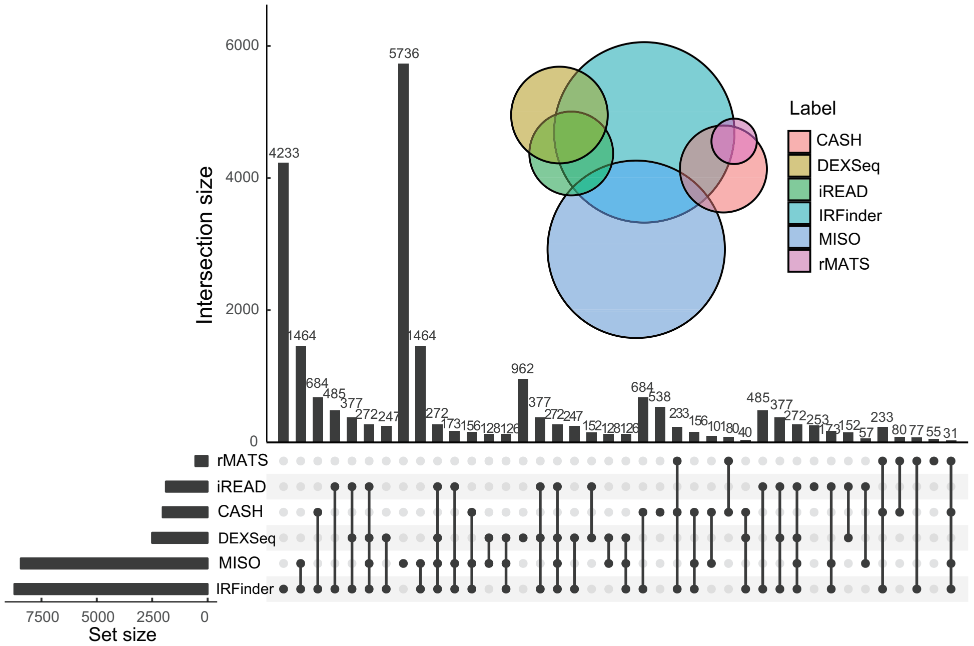


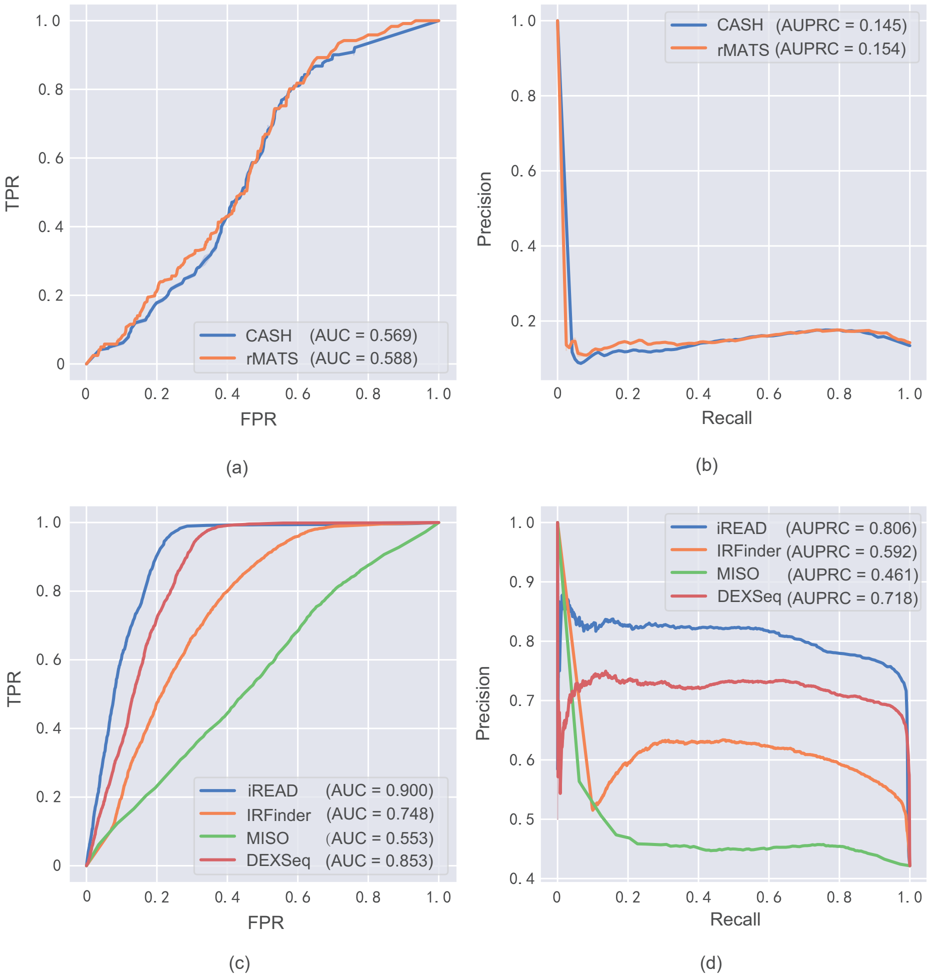
Fig. 1 **Venn and UpSet diagram showing the IR intersection of six methods on SIMU30. The Venn diagram is in the upper right corner. The color of the circle indicates the method and the size indicates the number of IRs the method predicts. The UpSet diagram is a combined bar chart at the bottom. The horizontal bar chart on the left shows the total number of IRs predicted by each method. The connection point graph and the vertical bar chart together show the IR intersection between methods. A single point in the connection point graph represents the number of unique IRs predicted by each method.**

these methods is so limited that the detected IR could not be fully used, and the performance of these methods could not be well compared. Therefore, we divided the six methods into two categories based on how IR is analyzed and reported in each method. The first category contains rMATS and CASH, which are designed to analyze differentially expressed IRs for case-control experiments including multiple samples. The remaining four (iREAD, IRFinder, MISO, and DEXSeq) belong to the second category, which detects IR from only a single sample. The methods in the second category often detect much more IRs than those in the first category. In this way, the different types of methods could be better compared. The number of intersections of IRs predicted by rMATS and CASH was 901, including 121 positive samples and 780 negative samples. The intersection number of IRs predicted by IRFinder, iREAD, MISO, and DEXSeq was 7867, including 3241 positive and 4446 negative samples. The AUC values of iREAD, DEXSeq, IRFinder, and MISO were 0.900, 0.853, 0.748, and 0.553, respectively, whereas those of rMATS and CASH were 0.588 and 0.569, respectively. The AUC and AUPRC obtained by all methods are shown in Fig. 2. iREAD demonstrated the best prediction performance among these four methods, and MISO performed less well. In simulation data generation, the introduction of a random base may affect the estimation of exon expression level distribution by MISO. In addition, the performance of IRFinder was not as good as that of iREAD, which was related to the IR detection tendency mentioned earlier. The simulation data showed that more IRs had relatively higher expression.

Finally, we investigated the influence of sequencing depth on the IR detection performance of each method. Except for IRFinder, the AUC of other methods decreased as the depth increased. The positive correlation of IRFinder reflects the stability of IR discrimination from the side. The relationship between AUC and depth is shown in Table 3.

3.2 Performance comparison on real-world RNA-Seq data

In this section, we compared the IR prediction performance of each method on two real-world RNA-Seq datasets. One is the Upf2 knockout mouse dataset without biological replicates, and the other is the APP mutant mouse dataset with eight biological replicates (see details in Section 2). Similarly, we extracted the intersection of IR sets detected by all methods based on the intron cluster. Then, we compared the performances



(a)

(b)

(c)

(d)

Fig. 2 Performance comparison for six methods in identifying IRs on SIMU30, (a) and (c) in terms of AUC, (b) and (d) in terms of AUPRC.

Table 3 Comparison of AUC on the simulated datasets with different sequencing depths.

Dataset	MISO	DEXSeq	iREAD	IRFinder	rMATS	CASH
SIMU15	0.554 09	0.872 51	0.921 11	0.73635	0.602 67	0.625 90
SIMU30	0.552 68	0.852 84	0.900 49	0.74815	0.587 66	0.568 95
SIMU60	0.559 67	0.857 03	0.895 61	0.786 05	0.586 79	0.587 21

of these methods on the intersection.

Here we show the results of the KO group in the Upf2 knockout mouse dataset. In the case of no biological replicates, MISO, IRFinder, DEXSeq, CASH, rMATS, and iREAD detected 17 381, 5694, 2047, 1892, 666, and 582 IR events, respectively. The numbers (and proportions) of unique IR events of all methods were 8443 (48%), 3373 (56%), 1164 (56%), 674 (36%), 118 (18%), and 98 (17%), respectively. The intersection of IR sets on the KO sample is shown in Fig. 3 and Table 4. The number of IRs and unique IRs detected by MISO considerably exceeded those detected by other methods. One of the possible reasons is that MISO considers some AS transcripts that only appear in specific cells or conditions compared with other methods. In addition, the proportion of unique IRs predicted by each method varied greatly, which partly reflects the ability of each method to detect novel IRs.

Here we show the results for the APP mutant mouse dataset that contains 16 samples. MISO, IRFinder, DEXSeq, CASH, rMATS, and iREAD detected 18 309, 29 665, 3697, 4032, 1717, and 3079 IR events, respectively. The numbers (and proportions) of unique IR prediction results of all methods were 7750 (42%), 17 504 (59%), 1176 (32%), 959 (24%), 105 (6%), and 447 (15%), respectively. The intersection of IR sets among these methods is shown in Fig. 4. For the remaining 15 samples, IRFinder predicted the highest number of IRs, and the IR sets predicted by other methods had intersections with IRFinder. To show that these methods can be applied to the human dataset, we conducted evaluation experiments on a paired-

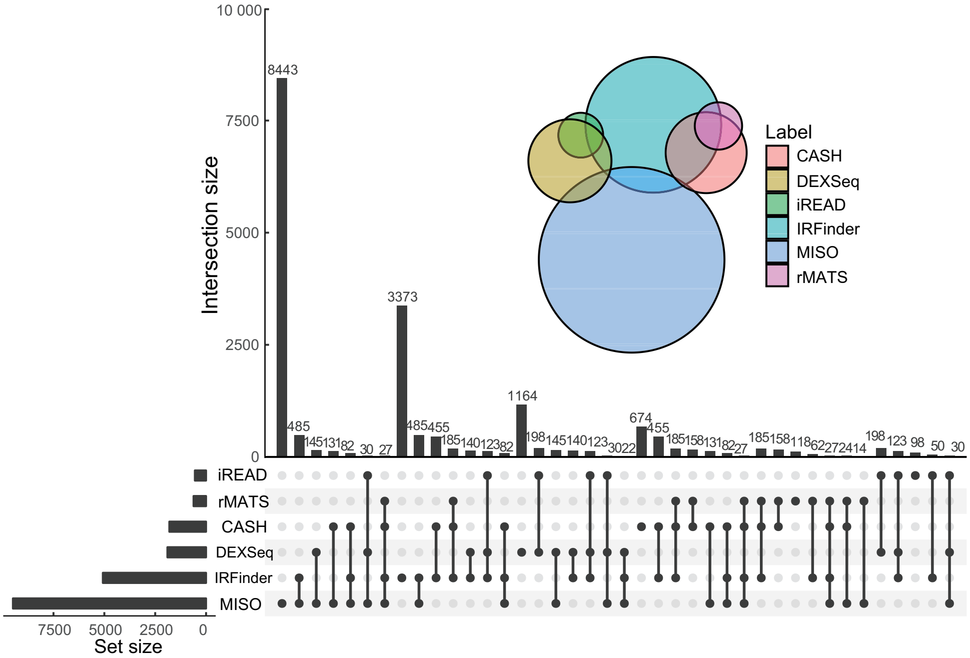


Fig. 3 Venn and UpSet diagram showing the IR intersection of six methods on the Upf2 knockout sample.

Table 4 The numbers of IRs detected on the Upf2 knockout mouse dataset.

Dataset	MISO	DEXSeq	IRFinder	iREAD	rMATS	CASH
Liver_WT	16 829	1624	6784	498	720	1878
Liver_KO	17 381	2047	5694	582	666	1892

end male dataset (accession ID in the SRA database: SRR5305480). The IR intersection results of each method are shown in Fig. A3 in the Appendix. MISO, IRFinder, DEXSeq, CASH, rMATS, and iREAD detected 8092, 2789, 1867, 1244, 535, and 130 IR events, respectively. The numbers (and proportions) of unique IR prediction results of all methods were 6971 (86%), 1544 (55%), 1390 (74%), 446 (36%), 127 (24%), and 19 (15%), respectively.

3.3 Differential analysis

These methods differ in their ways of detecting differentially-expressed IRs. The details are as

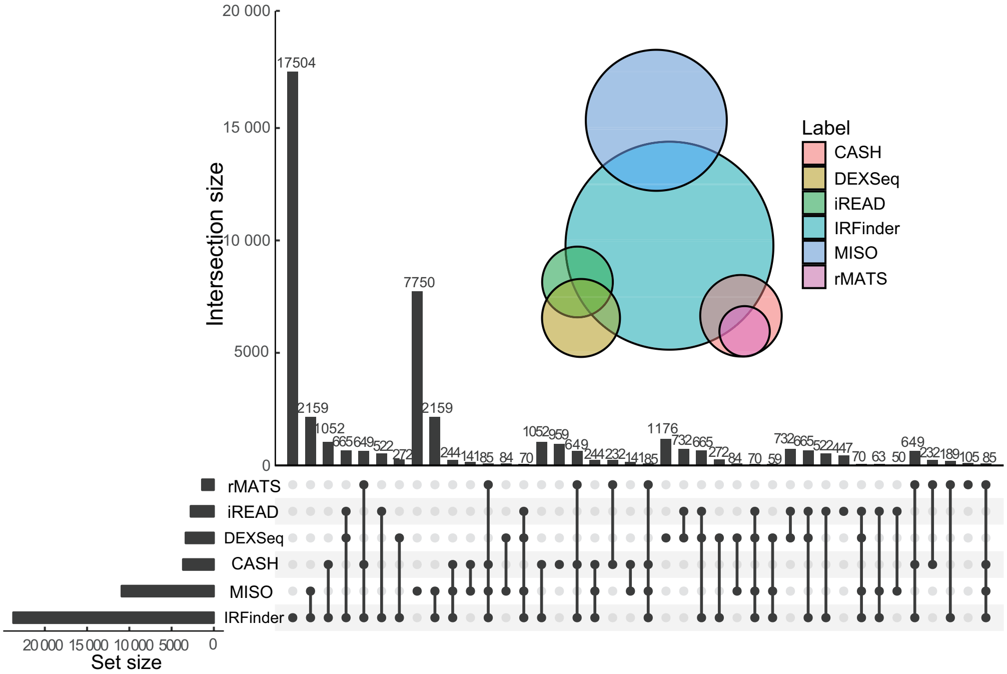


Fig. 4 Venn and UpSet diagram showing the IR intersection of six methods on the APP mutant mouse dataset.

follows. For iREAD, we combined the TR of all IR in each sample and then input them into edgeR (v3.26.8)[40] to detect significantly differentially expressed IRs. IRFinder uses read counts of introns to estimate the difference in read coverage between the experimental and control groups by fitting a GLM. DEXSeq provides an R script that encapsulates DESeq2 for differential analysis. The above three methods use log2FoldChange and FDR with appropriate thresholds ($|\text{log2FoldChange}| > 1$ and $\text{FDR}<0.05$) to detect differentially expressed IRs. rMATS uses delta-PSI ($|\text{delta-PSI}| > 0.05$) to represent the difference in intron inclusion level between groups and FDR ($\text{FDR}\leqslant 0.01$) to detect differentially expressed IRs. Because MISO is designed only for comparison between two samples, it cannot directly handle a control experiment with multiple biological replicates. One solution was to merge all BAM files by groups using SAMtools (v1.6)[35] and then input them into the pipeline of MISO for differential analysis. MISO uses the Bayesian factor ($\text{Bayesian factor}\geqslant 10$) and delta-PSI ($|\text{delta-PSI}| \geqslant 0.20$) to detect differentially expressed IRs.

First, we analyzed the number of differentially expressed IR detected by each method. For the differential analysis result of the APP mutant mouse dataset, the numbers of significantly differentially expressed IRs detected by MISO, IRFinder, rMATS, iREAD, DEXSeq, and CASH were 414, 361, 127, 41, 12, and 0, respectively. The intersection of differentially expressed IRs on the APP mutant dataset is shown in Fig. 5. MISO predicted the highest number of differentially expressed IRs. One possible reason was that we combined all the replicates in each group to accommodate the limitation that MISO can only compare two samples. DEXSeq uses a more general concept than differential AS, called DEU, which additionally considers changes in the usage of AS transcript polyadenylation sites and start sites. However, the number of differentially expressed IRs predicted by DEXSeq was relatively small. The reason may be that DEXSeq did not use the information of splice JR. In specific, CASH did not detect any differentially expressed IRs. The SpliceDiff module merged two statistical tests, and the conditions that required IR to pass both tests simultaneously were too strict.

Second, we mapped the Differentially Expressed IRs obtained by each method to Genes (DEIRG) and compared DEIRG with Differentially Expressed Genes (DEGs) identified using traditional exon-level expression using edgeR. We observed the difference between DEIRG and DEG. The number of traditional DEGs was 1031. The numbers of DEIRG obtained by MISO, IRFinder, rMATS, iREAD, and DEXSeq were 283, 103, 122, 15, and 16, respectively. The numbers of intersections between DEIRG and DEG for these

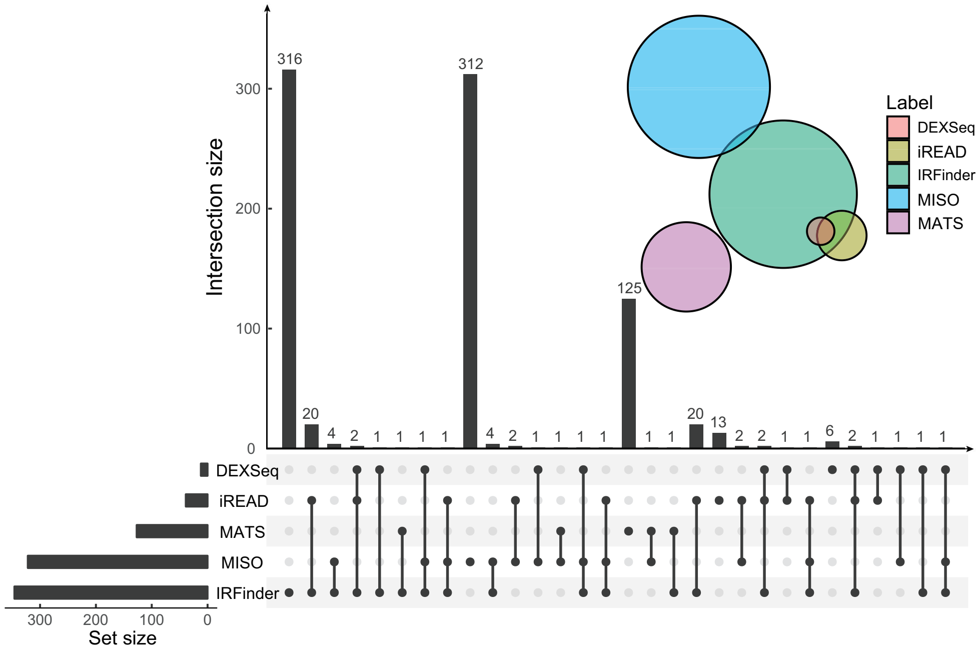


Fig. 5 Venn and UpSet diagram showing the comparison of significantly differentially expressed introns obtained by the six methods on the APP mutant mouse dataset.

methods were 35, 88, 6, 12, and 6, respectively. Many disease-related genes were detected by differentially expressed IRs. Among these DEIRGs, iREAD and IRFinder had the highest degree of overlap (over 80%). These results suggest that IR provides additional information for detecting disease-related genes that traditional exon-level based approaches cannot capture. Finally, we investigated the functions of DEIRGs by GO enrichment analysis with the R package clusterProfiler (v3.12)[41]. We set strict thresholds (Pvalue⩽0.01 and FDR⩽0.01) to detect enriched GO terms. The numbers of GO terms of each method and the intersection between methods are shown in Fig. 6. Only one intersection was found among IRFinder, DEXSeq, and iREAD, which was *astrocyte development* (GO: 0014002). This term was reported in the study of prefrontal gene expression patterns in rats by Duggan et al.[42] In addition, only one intersection was found between iREAD and DEXSeq, which is *neuron projection organization* (GO: 0106027). This term was reported in the study on the interaction between AD and the *Beta-secretase 2* gene[43]. The top 10 enriched GO terms of each method are shown in Fig. 7. For further verification, we searched the genes and gene products annotated by these GO terms and found that most of them contain genes that have been recorded in known AD-related gene databases[44–47].

4 Computational Time

Our experimental machine is a Linux system server with an Intel Xeon E5-2630 v4 10-core CPU and 252 GB memory. For the same reason, we separately compared the two categories of methods for the computational time. The results are shown in Fig. 8. rMATS and CASH were grouped into the first group to detect IR using two real-world control experiment datasets with no biological replicates: Upf2 knockout mouse dataset and APP mutant mouse dataset. rMATS can be set to run in parallel. The running time results showed that rMATS required much less running time than CASH. IRFinder, iREAD, DEXSeq, and MISO were grouped into the second group to detect IR using three simulation datasets with different sequencing depths: SIMU15, SIMU30, and SIMU60. Among these methods, iREAD and MISO can set parallel parameters, whereas IRFinder requires the input BAM file sorted by the read name. The running time results showed that iREAD and IRFinder took less time to run, whereas DEXSeq took the most.

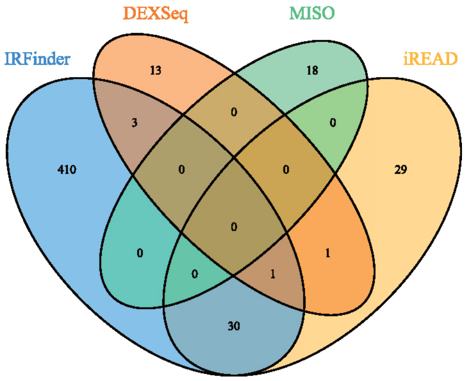


Fig. 6 Venn diagram showing the intersection of GO terms enriched in the retained introns obtained by IRFinder, DEXSeq, iREAD, and MISO. The only GO term shared by IRFinder, DEXSeq, and iREAD is the *astrocyte development* (GO: 0014002), whose annotated genes include the known AD-associated gene APP. The only GO term shared by iREAD and DEXSeq is the *neuron projection organization* (GO: 0106027), whose annotated gene includes the AD risk gene PRNP.

5 Discussion

In recent years, IR has received increasing attention because of its correlation with gene expression regulation and complex diseases[48–55]. Many researchers have made continuous efforts for the development of IR detection methods. These methods differ in many aspects, such as ways to preprocess RNA-Seq data, metrics to quantify the retention propensity, ways to detect differentially expressed IR, and so on. Although the principles of these methods have been reviewed[34], a systematic comparison of these methods has yet to be conducted.

In this article, we conducted a systematic experimental comparison of existing IR detection methods. First, considering that a gold standard of IR dataset is unavailable, we compared the IR detection performance on three simulation datasets for which the ground truth of retained introns is known. BEER was used to generate the simulated datasets with different sequencing depths. Second, the number of shared introns among more than three methods is small because of the different intron annotations and principles adopted by the methods. We used “intron cluster” to solve this problem, which is defined as connected intron regions within one gene as proposed in LeafCutter[37]. rMATS and CASH analyze differential AS events and require two groups of samples. Thus, these two methods were separately discussed from other ones.

Existing methods have limitations. First, in terms

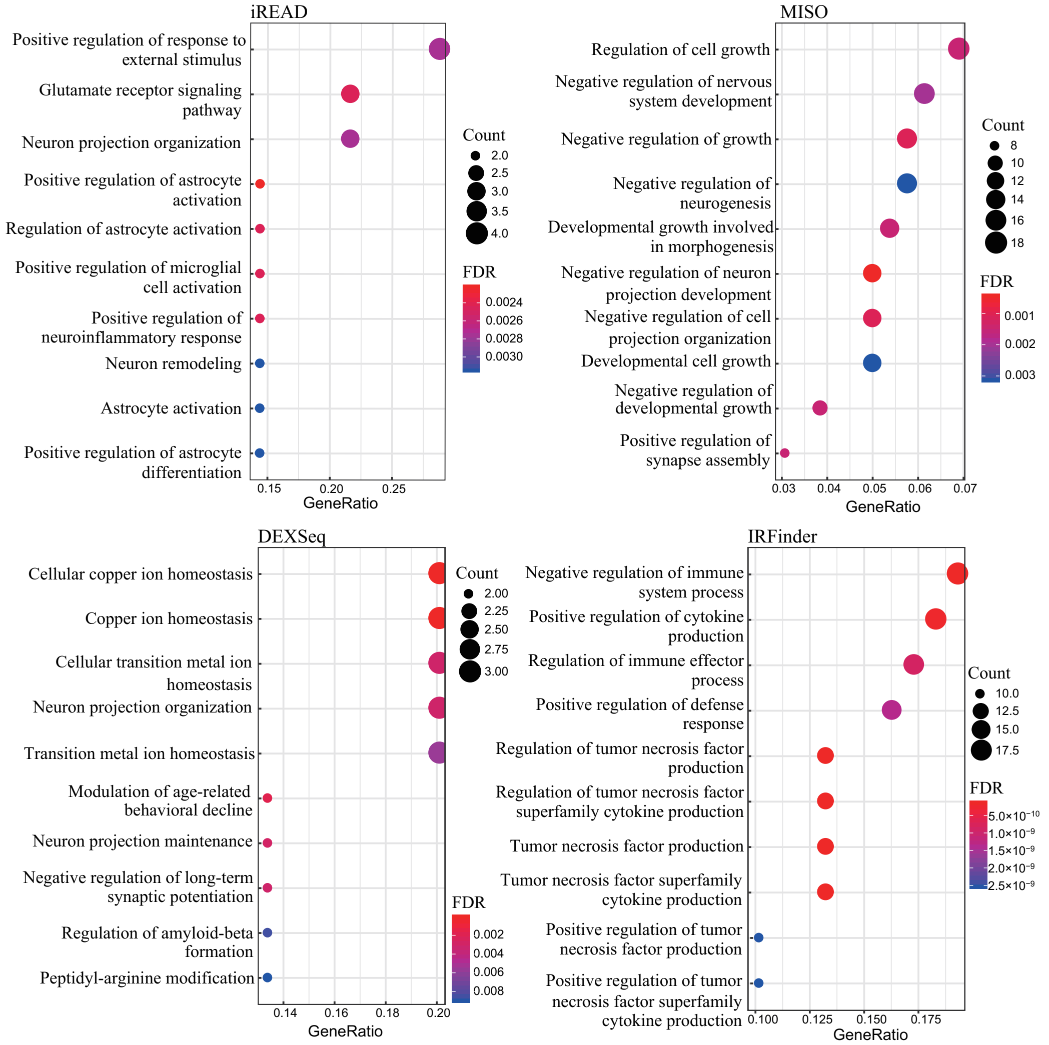
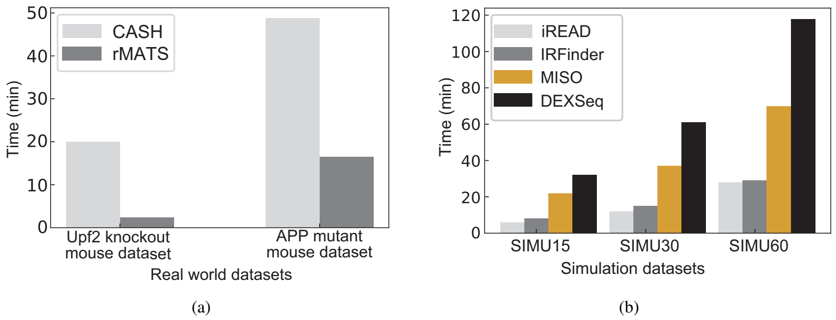


Fig. 7 Top 10 GO enrichment results for rMATS, IRFinder, iREAD, and DEXSeq, where GeneRatio indicates the proportion of genes containing the GO term. Circle size represents the number of genes or gene products annotated by GO terms. The more reddish the color, the more significant the enrichment.



(a) (b)

Fig. 8 Comparison of running time of IR detection methods. (a) Running time of CASH and rMATS on two real-world datasets without replicates. (b) Running time of IRFinder, iREAD, MISO, and DEXSeq on three simulated datasets with different sequencing depths.

of methodological integrity, IRFinder has the highest functional integration, which integrates sequence alignment, sample quality control, IR detection, and differential analysis. Each method can perform sample quality control before IR detection, such as whether or not DNA contamination is present, which helps reduce noise. Second, in terms of the information on which IR detection depends, relying solely on the annotated transcript could limit the discovery of novel IRs because the current transcript annotation is incomplete. The intron coordinates constructed by sequence alignment of RNA-Seq data can remove the dependence on incomplete annotation and are conducive to the detection of novel IRs. For example, CASH combines transcript annotation and RNA-Seq data information to reconstruct AS sites. Third, current methods detect IRs based on only RNA-Seq data without considering other information, such as sequence features and splice site strength[56–58]. The introduction of intron-related a priori information, such as intron length and canonical status of splice sites, is helpful for IR detection. Fourth, the discussed methods do not resolve IR events at the isoform level because of the short read length of RNA-Seq technology. That is, retained introns are not assigned to their origin of isoforms. IR can be detected at the isoform level in two possible ways. One is to construct intron retaining isoforms according to the detected IRs combined with transcriptome annotation and then quantify them using transcriptome quantitative tools. However, the construction of isoforms has a combinatorial explosion problem. The other is to use third-generation sequencing technologies, such as SMRT[59], which can generate full-length transcript isoforms. Long sequencing length and high sequencing error rate are the characteristics of third-generation sequencing. The long read length is the main reason why the current IR detection methods cannot be directly applied to third-generation sequencing data. Accordingly, the calculation method of coverage depth from sequencing data needs to be improved to expand the data types available for existing IR detection methods. In the future, developing methods to detect IRs for isoforms would be of great value. Addressing the above limitations could enable the development of accurate IR detection methods.

6 Conclusion

IR has been gaining increasing attention for its roles in gene regulation and its association with diseases. Analysis of IR provides a complementary approach to traditional methods that are based on only exon expression. IR is informative about disease status or biomarkers. In the analysis of IR, the first step is to detect IR events from biological samples. Several methods have been developed, each having advantages and limitations. This work discussed key principles and features of existing IR detection approaches and performed systematic experiments to compare their performance on detecting IRs. We pinpointed potential ways to improve the accuracy of IR detection methods. We expect that the systematic analysis provides a helpful guidance for the detection and analysis of IRs.

Appendix

Figures A1–A3 are presented in this section.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61772556, 61972185, U1909208, 61972423, and 61832019), 111 Project (No. B18059), and Hunan Provincial Science and

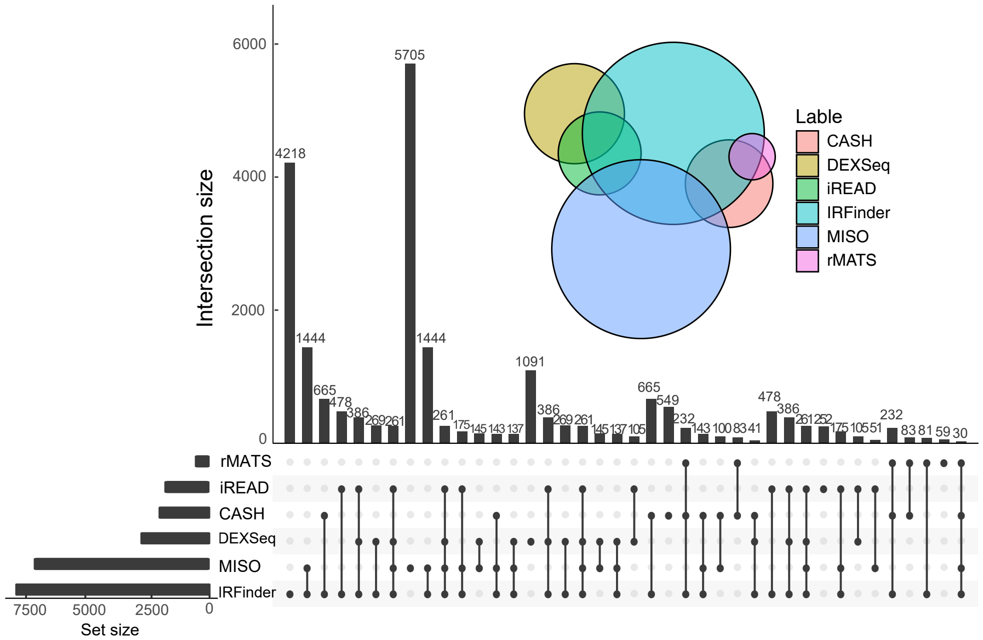


Fig. A1 A combination diagram of Venn and UpSet shows the IR intersection of six methods on SIMU15.

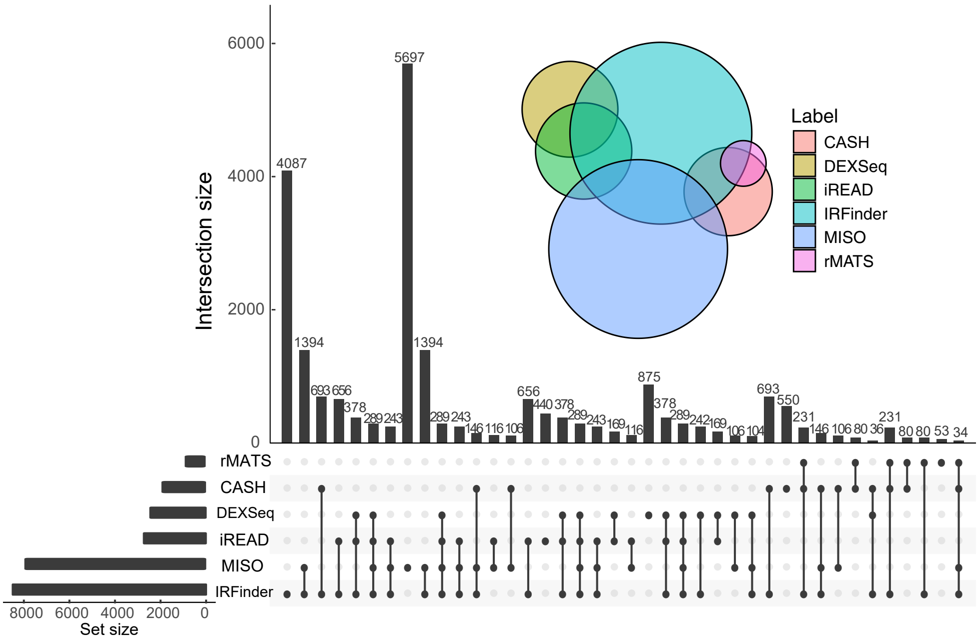


Fig. A2 A combination diagram of Venn and UpSet shows the IR intersection of six methods on SIMU60.

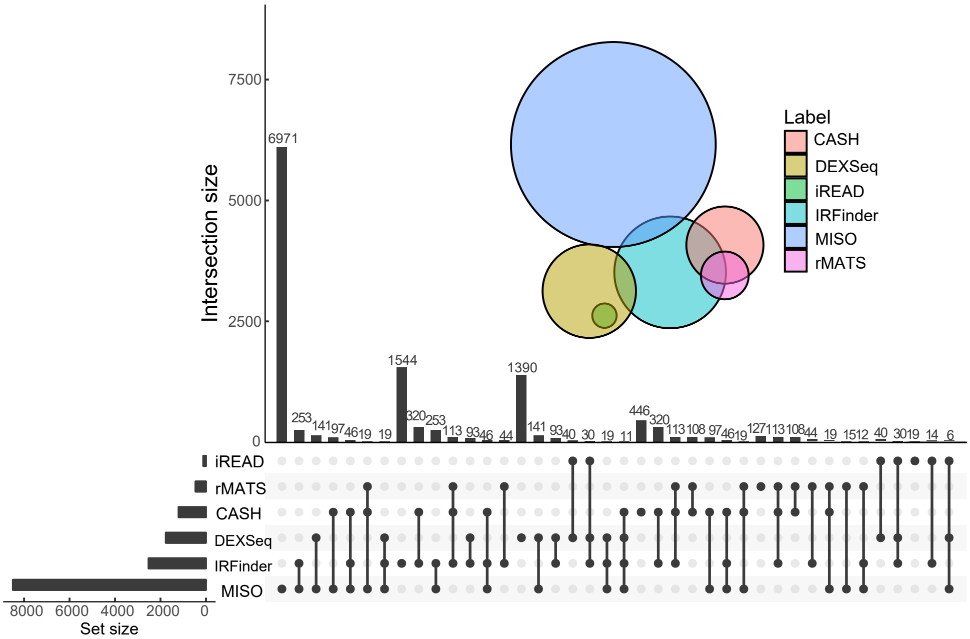


Fig. A3 A combination diagram of Venn and UpSet shows the IR intersection of six methods on human SRR5305480.

Technology Program (No. 2018WK4001).

The results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Support for these studies was provided by the NIH U01 AG046139. We thank Drs. Jada Lewis, Karen Duff, David Westaway, and David Borchelt for generating these lines of transgenic mice and providing us access to them.

References

[1] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz, Alternative splicing: A pivotal step between eukaryotic transcription and translation, *Nat. Rev. Mol. Cell Biol.*, vol. 14, no. 3, pp. 153–165, 2013.

[2] S. Chaudhary, W. Khokhar, I. Jabre, A. S. N. Reddy, L. J. Byrne, C. M. Wilson, and N. H. Syed, Alternative splicing and protein diversity: Plants versus animals, *Front. Plant Sci.*, vol. 10, p. 708, 2019.

[3] F. E. Baralle and J. Giudice, Alternative splicing as a regulator of development and tissue identity, *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 7, pp. 437–451, 2017.

[4] S. A. Bhuiyan, S. Ly, M. Phan, B. Huntington, E. Hogan, C. C. Liu, J. Liu, and P. Pavlidis, Systematic evaluation of

isoform function in literature reports of alternative splicing, *BMC Genomics*, vol. 19, no. 1, p. 37, 2018.

[5] G. Biamonti, A. Amato, E. Belloni, A. Di Matteo, L. Infantino, D. Pradella, and C. Ghigna, Alternative splicing in Alzheimer's disease, *Aging Clin. Exp. Res.*, vol. 33, no. 4, pp. 747–758, 2019.

[6] E. El Marabti and I. Younis, The cancer spliceome: Reprograming of alternative splicing in cancer, *Front. Mol. Biosci.*, vol. 5, p. 80, 2018.

[7] A. C. H. Wong, J. E. J. Rasko, and J. J. L. Wong, We skip to work: Alternative splicing in normal and malignant myelopoiesis, *Leukemia*, vol. 32, no. 5, pp. 1081–1093, 2018.

[8] A. Paschalis, A. Sharp, J. C. Welti, A. Neeb, G. V. Raj, J. Luo, S. R. Plymate, and J. S. De Bono, Alternative splicing in prostate cancer, *Nat. Rev. Clin. Oncol.*, vol. 15, no. 11, pp. 663–675, 2018.

[9] E. Fraile-Bethencourt, A. Valenzuela-Palomo, B. Díez-Gómez, E. Goina, A. Acedo, E. Buratti, and E. A. Velasco, Mis-splicing in breast cancer: Identification of pathogenic *BRCA2* variants by systematic minigene assays, *J. Pathol.*, vol. 248, no. 4, pp. 409–420, 2019.

[10] P. A. F. Galante, N. J. Sakabe, N. Kirschbaum-Slager, and S. J. De Souza, Detection and evaluation of intron retention events in the human transcriptome, *RNA*, vol. 10, no. 5, pp. 757–765, 2004.

[11] N. J. Sakabe and S. J. De Souza, Sequence features responsible for intron retention in human, *BMC Genomics*, vol. 8, no. 1, p. 59, 2007.

[12] R. Louro, A. S. Smirnova, and S. Verjovski-Almeida, Long intronic noncoding RNA transcription: Expression noise or expression choice? *Genomics*, vol. 93, no. 4, pp. 291–298, 2009.

[13] C. Cenik, A. Derti, J. C. Mellor, G. F. Berriz, and F. P. Roth, Genome-wide functional analysis of human 5' untranslated region introns, *Genome Biol.*, vol. 11, no. 3, p. R29, 2010.

[14] C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov, Selection for short introns in highly expressed genes, *Nat. Genet.*, vol. 31, no. 4, pp. 415–418, 2002.

[15] Q. Zhang, H. Li, H. Jin, H. B. Tan, J. Zhang, and S. T. Sheng, The global landscape of intron retentions in lung adenocarcinoma, *BMC Med. Genomics*, vol. 7, no. 1, p. 15, 2014.

[16] D. Wang, J. Zavadil, L. Martin, F. Parisi, E. Friedman, D. Levy, H. Harding, D. Ron, and L. B. Gardner, Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis, *Mol. Cell. Biol.*, vol. 31, no. 17, pp. 3670–3680, 2011.

[17] C. T. Ong and S. Adusumalli, Increased intron retention is linked to Alzheimer's disease, *Neural Regen. Res.*, vol. 15, no. 2, pp. 259–260, 2020.

[18] H. Jung, D. Lee, J. Lee, D. Park, Y. J. Kim, W. Y. Park, D. W. Hong, P. J. Park, and E. Lee, Intron retention is a widespread mechanism of tumor-suppressor inactivation, *Nat. Genet.*, vol. 47, no. 11, pp. 1242–1248, 2015.

[19] H. Dvinge and R. K. Bradley, Widespread intron retention diversifies most cancer transcriptomes, *Genome Med.*, vol. 7, no. 1, p. 45, 2015.

[20] S. R. Zhao, Alternative splicing, RNA-Seq and drug discovery, *Drug Discov. Today*, vol. 24, no. 6, pp. 1258–1267, 2019.

[21] J. Feng, K. Chen, X. Dong, X. L. Xu, Y. X. Jin, X. Y. Zhang, W. B. Chen, Y. J. Han, L. Shao, Y. Gao, et al., Genome-wide identification of cancer-specific alternative splicing in circRNA, *Mol. Cancer*, vol. 18, no. 1, p. 35, 2019.

[22] V. Van Giau, E. Bagyinszky, Y. S. Yang, Y. C. Youn, S. S. A. An, and S. Y. Kim, Genetic analyses of early-onset Alzheimer's disease using next generation sequencing, *Sci. Rep.*, vol. 9, no. 1, p. 8368, 2019.

[23] Y. Bai, S. F. Ji, and Y. D. Wang, IRcall and IRclassifier: Two methods for flexible detection of intron retention events from RNA-Seq data, *BMC Genomics*, vol. 16, no. 2, p. S9, 2015.

[24] H. Pimentel, J. G. Conboy, and L. Pachter, Keep me around: Intron retention detection and analysis, arXiv preprint arXiv: 1510.00696, 2015.

[25] A. Roberts and L. Pachter, Streaming fragment assignment for real-time analysis of sequencing experiments, *Nat. Methods*, vol. 10, no. 1, pp. 71–73, 2013.

[26] R. Middleton, D. D. Gao, A. Thomas, B. Singh, A. Au, J. J. L. Wong, A. Bomane, B. Cosson, E. Eyras, and J. E. J. Rasko, et al., IRFinder: Assessing the impact of intron retention on mammalian gene expression, *Genome Biol.*, vol. 18, no. 1, p. 51, 2017.

[27] H. D. Li, C. C. Funk, and N. D. Price, iREAD: A tool for intron retention detection from RNA-Seq data, *BMC Genomics*, vol. 21, no. 1, p. 128, 2020.

[28] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, Analysis and design of RNA sequencing experiments for identifying isoform regulation, *Nat. Methods*, vol. 7, no. 12, pp. 1009–1015, 2010.

[29] S. H. Shen, J. W. Park, J. Huang, K. A. Dittmar, Z. X. Lu, Q. Zhou, R. P. Carstens, and Y. Xing, MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data, *Nucleic Acids Res.*, vol. 40, no. 8, p. e61, 2012.

[30] S. H. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 51, pp. E5593–E5601, 2014.

[31] S. Anders, A. Reyes, and W. Huber, Detecting differential usage of exons from RNA-Seq data, *Nat. Prec.*, doi: 10.1038/npre.2012.6837.2.

[32] Y. F. Li, X. Y. Rao, W. W. Mattox, C. I. Amos, and B. Liu, RNA-Seq analysis of differential splice junction usage and intron retentions by DEXSeq, *PLoS One*, vol. 10, no. 9, p. e0136653, 2015.

[33] W. W. Wu, J. Zong, N. Wei, J. Cheng, X. X. Zhou, Y. M. Cheng, D. Chen, Q. H. Guo, B. Zhang, and Y. Feng, CASH: A constructing comprehensive splice site method for detecting alternative splicing events, *Brief. Bioinform.*, vol. 19, no. 5, pp. 905–917, 2018.

[34] L. Broseus and W. Ritchie, Challenges in detecting and quantifying intron retention from next generation

sequencing data, *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 501–508, 2020.

[35] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools, *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[36] G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch, and E. A. Pierce, Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM), *Bioinformatics*, vol. 27, no. 18, pp. 2518–2528, 2011.

[37] Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, and J. K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter, *Nat. Genetics*, vol. 50, pp. 151–158, 2018.

[38] H. D. Li, GTFtools: A Python package for analyzing various modes of gene models, *bioRxiv*, doi: 10.1101/263517.

[39] A. R. Quinlan and I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features, *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[40] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, edgeR: A bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[41] G. C. Yu, L. G. Wang, Y. Y. Han, and Q. Y. He, clusterProfiler: An R package for comparing biological themes among gene clusters, *OMICS: A J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, 2012.

[42] M. R. Duggan, S. Joshi, Y. F. Tan, M. Slifker, E. A. Ross, M. Wimmer, and V. Parikh, Transcriptomic changes in the prefrontal cortex of rats as a function of age and cognitive engagement, *Neurobiol. Learn. Mem.*, vol. 163, p. 107035, 2019.

[43] A. De Lillo, G. A. Pathak, F. De Angelis, M. Di Girolamo, M. Luigetti, M. Sabatelli, F. Perfetto, S. Frusconi, D. Manfellotto, M. Fuciarelli, et al., Epigenetic profiling of Italian patients identified methylation sites associated with hereditary transthyretin amyloidosis, *medRxiv*, doi: 10.1101/2020.04.13.20064006.

[44] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res.*, vol. 33, no. S1, pp. D514–D517, 2005.

[45] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Z. Huang, R. Lopez, et al., The Universal Protein resource (UniProt): An expanding universe of protein information, *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D187–D191, 2006.

[46] Z. X. Bai, G. C. Han, B. Xie, J. J. Wang, F. H. Song, X. Peng, and H. X. Lei, AlzBase: An integrative database for gene dysregulation in Alzheimer's disease, *Mol. Neurobiol.*, vol. 53, no. 1, pp. 310–319, 2016.

[47] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D833–D839, 2017.

[48] D. P. Vanichkina, U. Schmitz, J. J. L. Wong, and J. E. J. Rasko, Challenges in defining the role of intron retention in normal biology and disease, *Semin. Cell Dev. Biol.*, vol. 75, pp. 40–49, 2018.

[49] A. C. Smart, C. A. Margolis, H. Pimentel, M. X. He, D. A. Miao, D. Adeegbe, T. Fugmann, K. K. Wong, and E. M. Van Allen, Intron retention as a novel source of cancer neoantigen, *bioRxiv*, doi: 10.1101/309450.

[50] D. X. Zhang, Q. Hu, X. Z. Liu, Y. B. Ji, H. P. Chao, Y. Liu, A. Tracz, J. Kirk, S. Buonamici, and P. Zhu, et al., Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer, *Nat. Commun.*, vol. 11, no. 1, p. 2089, 2020.

[51] D. Kim, M. Shivakumar, S. Han, M. S. Sinclair, Y. J. Lee, Y. L. Zheng, O. I. Olopade, D. Kim, and Y. Lee, Population-dependent intron retention and DNA methylation in breast cancer, *Mol. Cancer Res.*, vol. 16, no. 3, pp. 461–469, 2018.

[52] H. D. Li, R. Menon, G. S. Omenn, and Y. F. Guan, The emerging era of genomic data integration for analyzing splice isoform function, *Trends Genet.*, vol. 30, no. 8, pp. 340–347, 2014.

[53] H. D. Li, C. H. Yang, Z. M. Zhang, M. Y. Yang, F. X. Wu, G. S. Omenn, and J. X. Wang, IsoResolve: Predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation, *Bioinformatics*, vol. 37, no. 4, pp. 522–530, 2021.

[54] R. Eksi, H. D. Li, R. Menon, Y. C. Wen, G. S. Omenn, M. Kretzler, and Y. F. Guan, Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-Seq data, *PLOS Comput. Biol.*, vol. 9, no. 11, p. e1003314, 2017.

[55] Z. Y. Fang, C. X. Lin, Y. P. Xu, H. D. Li, and Q. S. Xu, REBET: A method to determine the number of cell clusters based on batch effect removal, *Brief. Bioinform.*, doi: 10.1093/BIB/BBAB204.

[56] J. T. Zheng, C. X. Lin, Z. Y. Fang, and H. D. Li, Intron retention as a mode for RNA-Seq data analysis, *Front. Genet.*, vol. 11, p. 586, 2020.

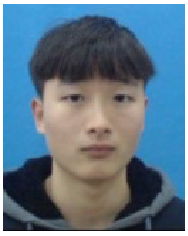
[57] A. Y. Zhang, S. A. Su, A. P. Ng, A. Z. Holik, M. L. Asselin-Labat, M. E. Ritchie, and C. W. Law, A data-driven approach to characterising intron signal in RNA-Seq data, *bioRxiv*, doi: 10.1101/352823.

[58] H. D. Li, C. C. Funk, K. McFarland, E. B. Dammer, M. Allen, M. M. Carrasquillo, Y. Levites, P. Chakrabarty, J. D. Burgess, and X. Wang, et al., Integrative functional genomic analysis of intron retention in human and mouse brain with Alzheimer's disease, *Alzheimer's Dementia*, vol. 17, no. 6, pp. 984–1004, 2021.

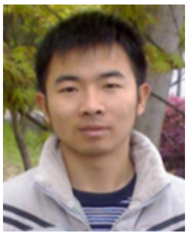
[59] D. An, H. X. Cao, C. S. Li, K. Humbeck, and W. Q. Wang, Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes, *Genes*, vol. 9, no. 1, p. 43, 2018.

Jiantao Zheng received the BEng degree from Zhengzhou University in 2018. He is currently a master student in bioinformatics at Central South University. He has published 1 SCI indexed research paper. His main research area is intron retention detection.

Cuixiang Lin received the BEng and MEng degrees both from Central South University in 2008 and 2010, respectively. She is currently a PhD candidate at Central South University. She has published 3 papers since 2018. Her research interests include identification of biomarkers of disease and disease mechanism.



Zhenpeng Wu received the BEng degree from Hunan Institute of Science and Technology in 2019. He is currently a master student in bioinformatics at Central South University. He has published 1 research paper in international conference. His main research area is intron retention detection.



Hong-Dong Li received the BEng (pharmaceutical engineering) and PhD (analytical chemistry) degrees from Central South University, Changsha, China in 2007 and 2012, respectively. He is currently working as an associate professor (tenure track) at School of Computer Science and Engineering, Central South University, Changsha, China. He has published over 60 SCI indexed papers in decent journals, including *Alzheimer's Dementia*, *Trends in Genetics*, *Brief Bioinform*, *PLOS Comput. Biol.*, *Bioinformatics*, *IEEE ACM T-Comput. Bi.*, *J. Proteome Res.*, *Proteomics*, *Nat. Genet.* (*co-author*), *Nat. Commu.* (*co-author*), *Chemom. Intell. Lab. Syst*. According to Google Scholar, his publications have received over 4500 citations with H-index=32. The maximum citation number of a single paper is over 800. One paper was selected into the ESI 1% collection. He co-authored a monograph in English (CRC Press, USA) and wrote two book chapters. His research interests include bioinformatics, chemometrics, and machine learning.