BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/10 pp115–126
Volume 6, Number 2, June 2023
DOI: 10.26599/BDMA.2022.9020023

# DeepRetention: A Deep Learning Approach for Intron Retention Detection

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

**Abstract:** As the least understood mode of alternative splicing, Intron Retention (IR) is emerging as an interesting area and has attracted more and more attention in the field of gene regulation and disease studies. Existing methods detect IR exclusively based on one or a few predefined metrics describing local or summarized characteristics of retained introns. These metrics are not able to describe the pattern of sequencing depth of intronic reads, which is an intuitive and informative characteristic of retained introns. We hypothesize that incorporating the distribution pattern of intronic reads will improve the accuracy of IR detection. Here we present DeepRetention, a novel approach for IR detection by modeling the pattern of sequencing depth of introns. Due to the lack of a gold standard dataset of IR, we first compare DeepRetention with two state-of-the-art methods, i.e. iREAD and IRFinder, on simulated RNA-seq datasets with retained introns. The results show that DeepRetention outperforms these two methods. Next, DeepRetention performs well when it is applied to third-generation long-read RNA-seq data, while IRFinder and iREAD are not applicable to detecting IR from the third-generation sequencing data. Further, we show that IRs predicted by DeepRetention are biologically meaningful on an RNA-seq dataset from Alzheimer's Disease (AD) samples. The differential IRs are found to be significantly associated with AD based on statistical evaluation of an AD-specific functional gene network. The parent genes of differential IRs are enriched in AD-related functions. In summary, DeepRetention detects IR from a new angle of view, providing a valuable tool for IR analysis.

**Key words:** Alternative Splicing (AS); Intron Retention (IR); intronic reads distribution pattern; RNA-seq

## 1 Introduction

Through Alternative Splicing (AS), a single gene can produce a variety of splicing isoforms whose sequences, structures, and functions are different[1]. AS is a frequent phenomenon in eukaryotes, dramatically increasing the biological diversity of transcripts and coding proteins[2, 3]. AS includes five major types of modes: (1) exon skipping, (2) mutually exclusive exons, (3) alternative 5' splice sites, (4) alternative 3' splice sites, and (5) Intron Retention (IR)[4]. As the least common type of AS, IR means that the intron is retained in the mature mRNA rather than being spliced out as usual[5].

Recently, an increasing number of studies have shown that IR played a specific role in gene expression regulation[6]. Braunschweig et al.[7] performed comprehensive IR analysis on the high coverage RNA-seq data of more than 40 different cells and tissue types from humans and mice, and found that increased IR down-regulated *Ssrp1* mRNA expression during neuronal differentiation. Using human and mouse strand-specific CD4+ T cell data, Ni et al.[8] found that 12% (185/1583) of the genes were regulated by IR level

• Zhenpeng Wu, Jiantao Zheng, Jiashu Liu, Cuixiang Lin, 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: zhenpeng@csu.edu.cn; cs_jtzheng@163.com; 214711048@csu.edu.cn; lincxcsu@csu.edu.cn; hongdong@csu.edu.cn.

† Zhenpeng Wu and Jiantao Zheng contribute equally to this paper.

* To whom correspondence should be addressed.
Manuscript received: 2022-03-11; revised: 2022-06-07; accepted: 2022-07-06

and highly enriched in the proteasome pathway that was critical for the proliferation of T cell and release of cytokine.

Moreover, IR has been proved to be associated with complex diseases[9, 10]. Using mass spectrometry immunopeptidome analysis, Smart et al.[11] found that retained intron neoepitopes were processed and presented on the surface of cancer cell lines. Zhang et al.[9] found that as prostate cancer progression, the differential IR events between the patient and the control samples were increasing, and IR was established as a hallmark of prostate cancer stemness and aggressiveness. In addition, we performed a genome-wide analysis of IR by integrating genetic, transcriptomic, and proteomic data from Alzheimer's Disease (AD) samples[12]. We identified the AD-specific intron expression quantitative trait loci and found that the intron expression of innate immune genes was significantly correlated with AD.

The rapid development of high-throughput sequencing technology has laid the foundation for IR detection[13]. Evaluating IR in the transcriptome by computational methods is currently an emerging research field. However, the source of intronic reads is highly heterogeneous, which hinders the detection of IR[14]. Existing IR detection methods, such as iREAD[15] and IRFinder[16], are exclusively based on one or a few predefined metrics describing local or summarized characteristics of intronic reads. IRs are detected based on predefined thresholds of these metrics. The flatter the distribution of intronic reads is, the higher the authenticity of intron retention is. The flatness is not equal to IR level, but flatness is one of the importance features of retained introns. Therefore, iREAD quantifies the intronic reads distribution flatness by calculating the normalized entropy score, and then combines with several other complementary metrics such as the number of junction reads to evaluate IR. The predefined thresholds for iREAD are as follows: # total reads $\geqslant$ 20, # junction reads $\geqslant$ 1, FPKM $\geqslant$ 3, and NE-score $\geqslant$ 0.9. IRFinder defines IRratio as the proportion of intronic reads from intron-retaining transcripts, which depends on the median depth of introns. IRratio can reduce the effect of alternative splicing exons on gene coverage, but may tend to underestimate the retention level of long introns. The predefined thresholds for IRFinder are as follows: IRratio $\geqslant$ 0.1, coverage $\geqslant$ 0.7, SpliceExact $\geqslant$ 5, intron annotated with "clean" and "anti-near", and intron with the mark "–" or "NonUniformIntronCover". Broseus and Ritchie[17] developed a new version of IRFinder, called S-IRFindeR, in which the intron abundance is adjusted to the minimum coverage, providing a more rigorous but reliable estimation of IRratio. The metrics of these methods are not able to describe the pattern that how intronic reads are distributed (the pattern of sequencing depth), though the pattern is an intuitive and informative characteristic for IR detection. However, the distribution pattern of intronic reads has not been considered for IR detection.

Here we propose DeepRetention, a deep learning model that can capture the distribution patterns of intronic reads and use it for IR detection. DeepRetention extracts the fixed-length intron sequencing depth from multiple sub-segments of the intron, and then uses one-dimensional convolution filters to extract the pattern of sequencing depth. Using simulated RNA-seq data, we compare the performance of DeepRetention with state-of-the-art methods. DeepRetention is further verified to achieve high accuracy on the third-generation long-read RNA-seq data. We explore the characteristics of retained introns predicted by DeepRetention. Further, we show that DeepRetention is capable of predicting biologically meaningful IR events when applied to an RNA-seq dataset of AD samples. We test whether the differential IRs are significantly associated with AD based on statistical analysis. Using functional enrichment analysis, we illustrate that the parent genes of differential IRs are enriched in AD-related functions.

This paper is composed of four sections, and the contents of each section are discussed as follows: Section 1 introduces the research background and significance of IR, and the research status of IR detection. Section 2 describes the core idea of our DeepRetention model in detail, and then introduces the data processing. In Section 3, we compare DeepRetention with the existing methods, and explore the characteristics of retained introns predicted by DeepRetention. Section 4 summarizes the main contents of this paper and puts forward the prospect for the follow-up research.

## 2 Method and Material

### 2.1 DeepRetention

Existing methods detect IR exclusively based on one or several predefined metrics that describe the local or summarized characteristics of intronic reads. The metrics of existing methods are not able to capture the pattern of sequencing depth, which is an intuitive and

informative characteristic for IR detection. Therefore, in this work, we develop DeepRetention to predict IR by modeling the pattern of sequencing depth in intron regions. In addition, DeepRetention attempts to explore sequence information related to IR as auxiliary features, such as intron length. Therefore, DeepRetention takes the profile of intron sequencing depth as the main input. The intron sequence and intron length are also used as input. Finally, DeepRetention outputs the probability of introns being retained. Figure 1 illustrates the workflow of DeepRetention. DeepRetention mainly includes three parts: calculation of intron sequencing depth profiles, feature extraction, and IR prediction.

**(1) Calculation of intron sequencing depth profiles**

DeepRetention considers the sequencing depth profile as the pattern for IR detection. We uniformly divide the intron into $s (s \in [2, 3, \ldots, 10])$ different sub-segments. DeepRetention takes the raw sequencing depth of $s$ sub-segments as input. Specifically, DeepRetention uses the sequencing depth of 40 bases[18] on each side of the center point of the selected sub-segment as input. Therefore, the sequencing depth profile is a vector, which is composed of the read depth around each sub-segment and the total length of the profile is $\sum_{i=1}^{s}(40 + 40 + 1)$. The raw sequencing depth profile is extracted from RNA-seq data using "Samtools depth"[19]. The allowed minimum value of the sub-segments is 2, because the sequencing depth of 5' and 3' splice sites of the intron must be used as input for DeepRetention, which provides direct evidence to support intron retention[15]. The code is src/getReadCoverageSeq.py on GitHub (https://github.com/genemine/DeepRetention), users can retrain DeepRetention on their own RNA-seq data (see Subsection 2.3, model development).

**(2) Feature extraction**

DeepRetention requires three types of features for IR detection, including the distribution pattern of intron sequencing depth, intron sequence, and intron length. First, three sets of one-dimensional Convolution Blocks (CBs) are used to extract the vectorized representation of the distribution pattern of sequencing depth. CB is composed of two one-dimensional convolutional layers and a pooling layer. The number of convolution filters in CB gradually increases $(4 \rightarrow 16 \rightarrow 64)$[20], while the size of the convolution filter gradually decreases $(16 \rightarrow 8 \rightarrow 1)$. The convolution layer is filled with the "same padding"[21] way to ensure the consistency of dimensions. The pooling layer of the first two CBs is the maximum pooling, and the last CB is the global average pooling layer. The function of maximum pooling is

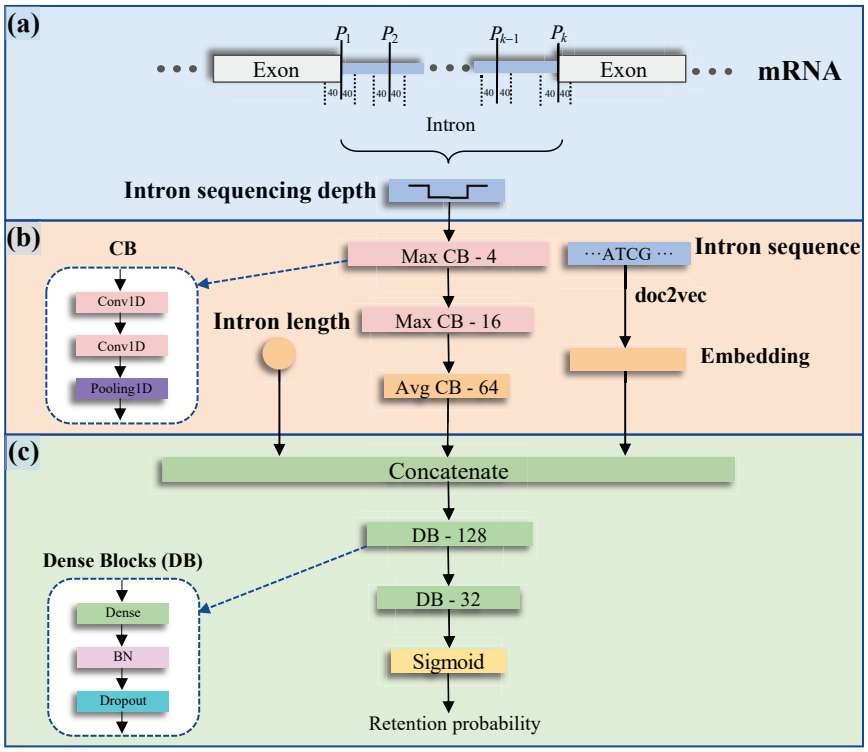



**Fig. 1 Workflow of DeepRetention. (a) Calculation of raw sequencing depth of introns to obtain the distribution pattern. (b) Extraction of the distribution pattern of intron sequencing depth, intron sequence, and intron length features. (c) DeepRetention integrates the above features, and outputs the probability of introns being retained.**

to extract the main features. The function of average pooling is to combine the features extracted by each convolution filter, and finally obtain the vectorized representation of the read depth distribution of introns. Second, the intron sequence is the nucleotide sequence of introns extracted from genome annotation. Then, doc2vec[22] is used to generate unified embedding representations of intron sequences. Specific steps are as follows: (1) The entire genome sequence is used for doc2vec training. According to gene coordinates, the gene sequence is extracted from the genome sequence. A gene sequence represents a train sample and is divided into $k$-mers ($k = 3$). For example, an AGCTA sequence is processed into [AGC, GCT, CTA]. The above-processed samples are used to train doc2vec and the vector size of doc2vec is specified as a fixed value (vector_size =100) during training. (2) The trained doc2vec is used to predict the sequence of each intron and will generate embedding vectors with fixed-length of 100. According to intron coordinates, the corresponding intron sequence is extracted from the genome sequence. The intron sequence is divided into $k$-mers and $k$-mers are input into doc2vec. The embedding vector of intron sequence is fixed-length vector because the vector size of doc2vec is specified as a fixed value during training. The application code of doc2vec is src/getReadSeq.py, and the training code is src/doc2vec.py on GitHub. Third, intron length is calculated based on the independent intron model, in which the coordinates of independent intron do not overlap with exons.

**(3) IR prediction**

DeepRetention concatenates the above features and passes them to two sets of DB, and finally outputs the probability of introns being retained. Each DB contains a fully connected layer, a Batch Normalization (BN) layer, and a dropout layer. Among them, the fully connected layer performs multiple sets of linear combinations on the output of the previous layer. The BN layer then standardizes them to promote the accelerated convergence of DeepRetention. The dropout layer randomly drops neurons with a given probability to prevent DeepRetention from overfitting. The final output layer is a fully connected layer with one neuron and sigmoid activation function. Sigmoid outputs a probability value ranging from 0 to 1, which represents the probability of intron being retained.

The DeepRetention models are implemented using Keras version 2.6.0. DeepRetention is freely available at https://github.com/genemine/DeepRetention.

## 2.2 RNA-seq data and processing

### 2.2.1 Simulated RNA-seq data

We use simulated RNA-seq data for comparison. The simulated RNA-seq data of 30 million paired reads (SIMU30) are obtained from Ref. [15]. The simulated data are generated by the BEER software[23] based on the mouse genome. The summary information of the dataset is shown in Table 1.

### 2.2.2 Real-world RNA-seq data

We obtained two real RNA-seq datasets for constructing training datasets (GM12878S: human; APPPS1: mouse; see Table 1). Due to the lack of a gold standard dataset of IR, we construct training datasets based on the union of the prediction results of iREAD (v0.8.9) and IRFinder (v1.3.1). After obtaining the intron intersection, the final label of each dataset is determined jointly by iREAD and IRFinder. Considering that iREAD and IRFinder are complementary in detecting IRs[15], if the sample is labeled as positive by one of iREAD or IRFinder, the intron is labeled as positive (retained introns). After determining all the positive introns, we sample the same number of negatives (spliced introns) as that of positives to construct the final training dataset. We keep the proportion balance of positive and negative samples in the training dataset. Table 2 shows the number of positive and negative samples in the training set.

Then, we obtain third-generation long-read RNA-seq data (GM12878T) for the validation of DeepRetention[25]. It contains over 10 million reads and is from the same cell line as GM12878S. The reads are aligned to the human reference genome (GRCh38) using Minimap2. We refer to the "naive" method introduced

**Table 1 Summary information of datasets used in this paper.**

| Dataset | Read length | Sequencing depth | Source |
|---|---|---|---|
| SIMU30 | 100 bp | 30 X | Reference [15] |
| GM12878S | 76 bp | 30 X | GSM958728 in Gene Expression Omnibus (GEO) |
| BBUKY | 101 bp | 70 X | Reference [24] |
| APPPS1 | 101 bp | 100 X | syn17008852 from Synapse |
| GM12878T | Full-length transcripts | 10 X | Reference [25] |

Note: bp denotes base pair.

**Table 2 Number of positive and negative samples in the training set.**

| Dataset | All | Positive | iREAD-positive | IRFinder-positive |
|---|---|---|---|---|
| GM12878S | 37 978 | 18 989 | 8299 | 13 374 |
| APPPS1 | 370 856 | 185 428 | 40 245 | 161 237 |

in S-IRFindeR[17] to identify IR from GM12878T. The number of long-reads spanning the intron is recorded as overall read abundance. The number of long-reads containing the intron is recorded as intronic read abundance. The proportion of intronic read abundance and overall read abundance is recorded as a ratio. In the subsequent analysis, overall read abundance $\geqslant 30$, intronic read abundance $\geqslant 20$, and ratio $\geqslant 0.05$ are used to filter unreliable estimates of ratio[17]. Then, if filtered introns are labeled IR by genome annotation, these introns will be retained. Finally, we obtain 804 retained introns.

We store the processed dataset in Zenodo (https://zenodo.org/record/6526078) for readers to download.

### 2.3 Model development

We construct two models based on the RNA-seq data from two species (GM12878S: human; APPPS1: mouse). The models are trained for 1000 epochs with a batch size of 128. A binary cross-entropy loss between the label and predicted output is minimized with Adam optimizer during training. The initial learning rate of Adam optimizer is set to 0.1; then it is reduced by a factor of 0.5 if the *val_loss* is not reduced in 5 epochs. Moreover, we repeat the training procedure 10 times and obtain 10 trained models, that is, 90% of data are used for each training, and the other 10% of data are used for validation. During testing, we evaluate the test set using all 10 trained models and their Area Under the Precision Recall Curve (AUPRC) and Area Under the Receiver Operating Characteristic curve (AUROC) scores are averaged as the final result. In detail, we will calculate AUPRC and AUROC of each model, and finally take the mean value of AUPRC and AUROC of 10 models.

## 3 Result

### 3.1 Influence of the number of intron sub-segments on DeepRetention model

We study the effects of the number of intron sub-segments on model performance. Each intron is evenly divided into $s$ sub-segments ($s \in [2, 3, \ldots, 10]$). The intron sequencing depth of $s$ sub-segments is input into the model. The test results of DeepRetention models with different $s$ values are shown in Fig. 2. We find that when $s$ is less than 7, the model performance increases with the increase of the $s$ value. When $s$ is greater than 7, the performance of the model does not change significantly. Therefore, we choose $s = 7$ as the optimal value, that is, we will take 7 different sub-segments uniformly in the intron region as the input of intron sequencing depth of the model.

### 3.2 Comparison between iREAD, IRFinder, and DeepRetention on simulated datasets

We evaluate iREAD, IRFinder, and DeepRetention on the simulated dataset. The three methods are comprehensively compared based on two metrics, including AUPRC and AUROC. Especially, we sort iREAD and IRFinder by normalized entropy score and IRratio, respectively. Table 3 shows an improvement of DeepRetention compared to iREAD and IRFinder. This is because DeepRetention detects IR by using the pattern

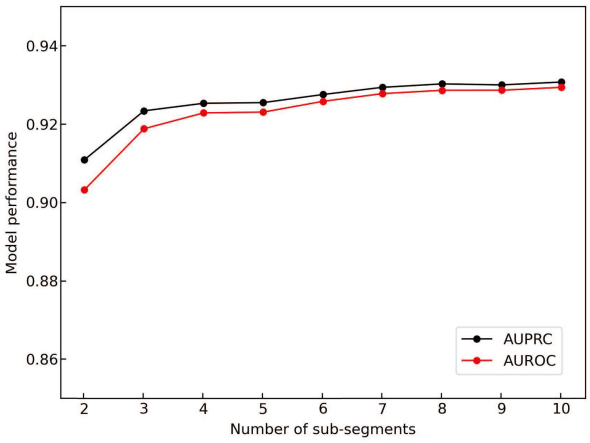



**Fig. 2 DeepRetention model performance varies with the number of intron sub-segments on the SIMU30 dataset. When $s = 2$, it means that DeepRetention takes the sequencing depth around the 5' and 3' splice sites of the intron. When $s \geqslant 3$, the intron is evenly divided into $s$ sub-segments, and the sequencing depth around each sub-segment will be input DeepRetention. Specifically, DeepRetention uses the sequencing depth of 40 bases[18] on each side of the center point of the selected sub-segment as input. Moreover, all sequencing depth of each segment will be used and the total length of sequencing depth is $\sum_{i=1}^{s} (40 + 40 + 1)$.**

**Table 3 Comparison results of iREAD, IRFinder, and DeepRetention on the simulated test dataset (SIMU30).**

| Method | AUPRC | AUROC |
|---|---|---|
| DeepRetention | 0.9293 | 0.9278 |
| iREAD | 0.8766 | 0.9004 |
| IRFinder | 0.8919 | 0.9206 |

of sequencing depth, which is similar to the visual result in Integrative Genomics Viewer (IGV) (see Subsection 3.4, characteristics analysis of retained introns).

### 3.3 Contribution of input features

To explore the contribution of different types of input features on the model performance, we only input a single type of feature to DeepRetention at a time (Table 4) and compare their corresponding performance. Because the simulated data are generated randomly, the sequence feature of retained introns cannot be modeled. The genome of real dataset can be used to characterize the sequence feature of retained introns. However, the next-generation sequencing lacks a gold standard dataset of IR. Therefore in this analysis, we use the third-generation sequencing dataset GM12878T, in which the retained introns are annotated by genome annotation[17]. Moreover, IRFinder and iREAD cannot detect IR from the third-generation sequencing dataset, so GM12878T is only used to explore individual input features. GM12878T has only 804 retained introns. To prevent class-imbalanced caused by few positive samples (retained introns), we randomly sample the same number of negative samples (non-retained introns) as that of positives. The sampling process is repeated 10 times to construct 10 sets of class-balanced validation datasets. The results of DeepRetention performance are shown in the form of mean plus/minus Standard Deviation (mean$\pm$SD).

The results show that DeepRetention is able to achieve high prediction accuracy and stability. DeepRetention achieves a stable and high AUPRC of 0.9420$\pm$0.0120 and AUROC of 0.9383$\pm$0.0082 (Table 4). Among the models with a single type of feature, the best performance is achieved when intron sequencing depth is used as the feature. The performance of the model with intron sequence or intron length as input is not significantly different, but much smaller than the model with all features and the model with intron sequencing depth only. This indicates that among the three input features, intron sequencing depth contributes most to the model performance, and intron sequence or intron length also contributes to IR detection.

**Table 4 Contribution analysis of a single type of feature to IR predictive performance based on the third-generation RNA-seq data.**

| Input feature | AUPRC | AUROC |
|---|---|---|
| All input features | 0.9420$\pm$0.0120 | 0.9383$\pm$0.0082 |
| Intron sequencing depth only | 0.9342$\pm$0.0144 | 0.9327$\pm$0.0055 |
| Intron sequence only | 0.6074$\pm$0.0094 | 0.6404$\pm$0.0105 |
| Intron length only | 0.6523$\pm$0.0115 | 0.6894$\pm$0.0079 |

### 3.4 Characteristics analysis of retained introns

We explore the characteristics of retained introns. The previous researches report that the retained introns have significantly weaker splice site strength[7, 26]. The lower the splicing strength is, the lower probability an intron will be spliced by the splicing factor. Therefore, we characterize the splice site strength of DeepRetention, iREAD and IRFinder based on the genome of real dataset. Overall, the results show that the splice site strength of all three methods is comparable, and the $p$-value of DeepRetention is more significant (Figures 3a and 3b). We find that no matter which prediction method is used, the splice site strength of retained introns is significantly lower than that of non-retained introns (see Figs. 3a and 3b; $p$-value $< 0.05$). For each method, the splice site strength between two types of introns is compared using the Mann-Whitney U test[28].

The retained introns have significantly higher Guanine-Cytosine (GC) content than non-retained introns[12], indicating that IR may be driven by the potential DNA sequences[29]. Therefore, we characterize the GC content of DeepRetention, iREAD, and IRFinder based on the genome of real dataset. Overall, the results show that the GC content of all three methods is comparable, and the $p$-values of DeepRetention and iREAD are more significant. Compared with non-retained introns, the retained introns have significantly higher GC content (see Fig. 3c; $p$-value $< 1.0 \times 10^{-6}$). For each method, the GC content between two types of introns is compared using the Mann-Whitney U test.

The retained introns may be regulated by RNA-Binding Proteins (RBPs)[30]. We perform motif analysis of human RBPs collected from Ref. [31] and the RBPmap database[32]. After removing the redundancy, we obtain 93 RBPs with known binding sites. We calculate the RBP-binding score within retained introns and non-retained introns using the same calculation method as the previous study[9]. Using the Mann-Whitney U test, we test whether the RBP-binding score of retained introns is significantly different from that of non-retained introns. The result shows that the RBP-binding scores of all three methods are comparable (see Fig. 3d), and the $p$-value of DeepRetention is more significant. Interestingly, we find that the RBP-binding score of retained introns is significantly lower than that of non-retained introns (see Fig. 3d; $p$-value

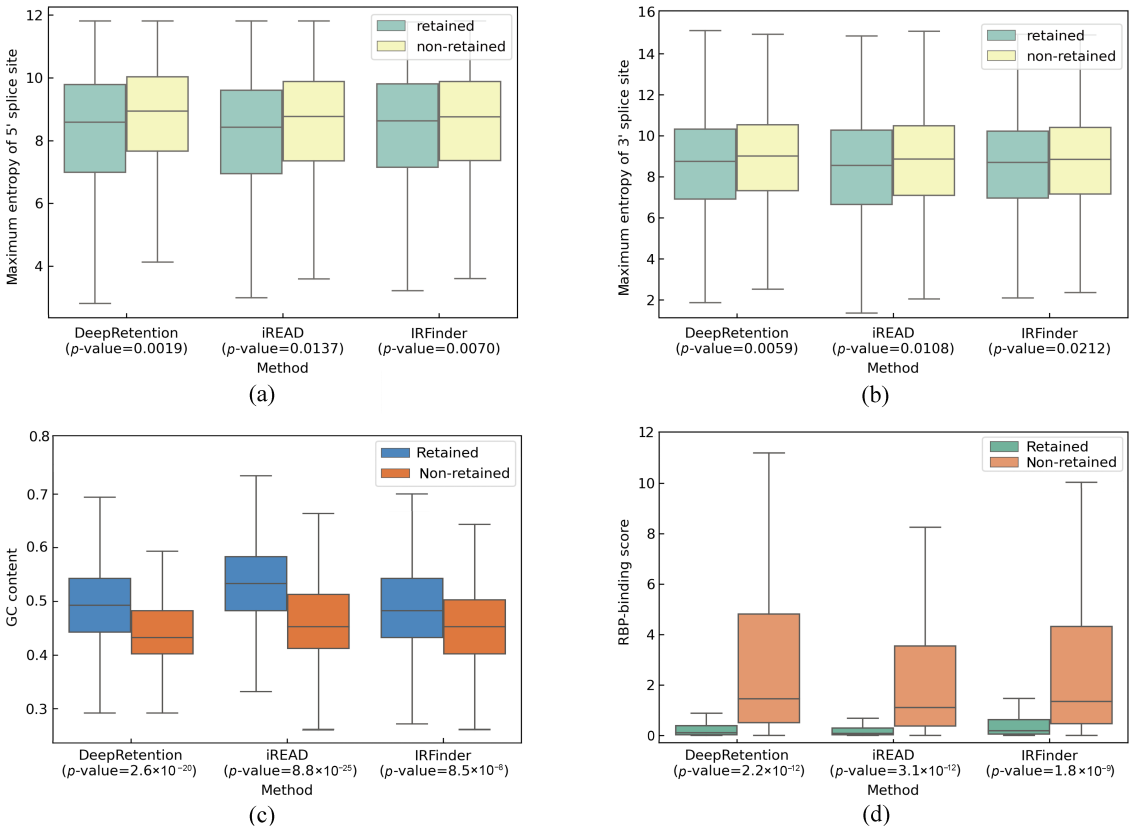



**Fig. 3 Characteristics analysis of the retained introns obtained by different methods. (a) The boxplot shows the splicing strength of 5' splice sites of retained introns and non-retained introns on the APPPS1 dataset. (b) The boxplot shows the splicing strength of 3' splice sites of retained introns and non-retained introns on the APPPS1 dataset. The splicing strength is calculated by MaxEntScan[27], which is based on the maximum entropy principle. (c) The boxplot shows the GC content of retained introns and non-retained introns on the APPPS1 dataset. (d) The boxplot shows the RBP-binding score within retained introns and non-retained introns on the GM12878S dataset. Since the collected RBPs are human RBPs, GM12878S is used for RBP-binding score calculation. The calculation method of the RBP-binding score is the same as the previous study[9]. Using the Mann-Whitney U test, the splicing strength, GC content, and RBP-binding score comparison between the retained and non-retained introns is made. The APPPS1 dataset is predicted by the model trained on GM12878S dataset, and the GM12878S dataset is predicted by the model trained on APPPS1 dataset.**

$< 1.0 \times 10^{-6}$), which is not uncovered by previous studies. This is consistent with our expectation, because the retention and splicing process of introns involves RBP, which belongs to one of the splicing factors mentioned above.

Then, we visually verify the retained introns only detected by our method (namely, DeepRetention-specific IR). We use the IGV software (v2.11.4)[33,34] to verify DeepRetention-specific IR. As illustrations, we show four retained introns in Fig. 4. In Fig. 4a, we show an retained intron (chr11: 57, 499, 805—57, 505, 233 of the gene *Fam114a2*) with a prediction score of 0.682. The second example is a confidently retained intron (chr3: 153, 909, 598—153, 910, 260 of the gene *Rabggtb*) with a prediction score of 0.905. In particular, the sequencing depth of some positions of this intron exceeds 30. These examples show that DeepRetention learned the pattern of sequencing depth. Therefore, our method can detect IR that cannot be detected by other methods.

### 3.5 Biological relevance of intron retention events predicted by DeepRetention

We test whether DeepRetention is capable of predicting biologically meaningful IR events. AD is a degenerative brain disease[35,36]. IR has been proved to be associated with AD[12]. We obtain RNA-seq data in fusiform gyrus of AD and control subjects[24], and these subjects derived from the Brain Bank of the University of KentuckY (BBUKY). We apply DeepRetention to detect IR on the BBUKY dataset. In this section, we consider that the intron is retained only if the prediction score is greater than 0.8 (0.5 by default in the previous sections). We first

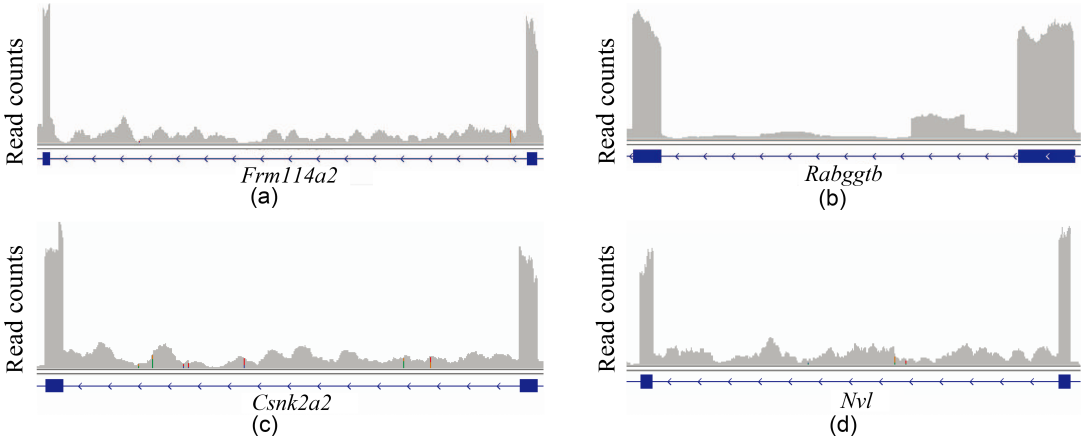
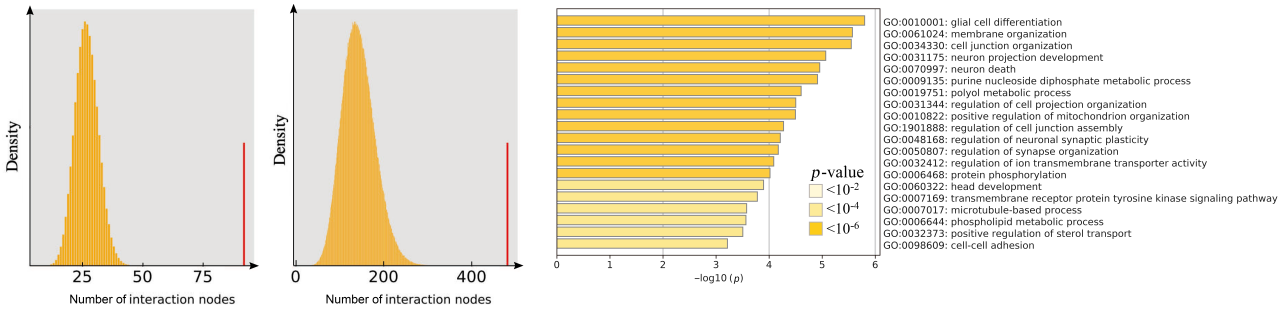



**Fig. 4 Illustration of examples of retained introns detected by DeepRetention on the APPPS1 dataset. These retained introns are not detected by iREAD and IRFinder. (a) The retained intron (chr11: 57, 499, 805—57, 505, 233) of *Fam114a2*. (b) The retained intron (chr3: 153, 909, 598—153, 910, 260) of *Rabggtb*. (c) The retained intron (chr8: 95, 457, 488—95, 460, 071) of *Csnk2a2*. (d) The retained intron (chr1: 181, 101, 635—181, 105, 071) of *Nvl*. The short blue box represents the exon and the blue line represents the intron.**

perform differential expression analysis on the BBUKY dataset using edgeR[37]. Because the input of edgeR is read count data, we take the read count of introns from iREAD outputs. We obtain 301 differential IRs between AD and control subjects (false discovery rate $< 0.05$ and fold change $> 2$). Then, we evaluate whether these differential IRs have biological significance based on two approaches, including statistical assessment with functional gene network and functional enrichment analysis.

In statistical assessment, we evaluate the parent genes of differential IRs based on our previous work on ADBrainNexus[38]. ADBrainNexus is a human AD-specific brain functional gene network constructed with AD genomic data. We used the collected 147 known AD-associated genes derived from multiple databases[38]. The association between the parent genes of differential IRs and the known AD-associated genes is tested based on statistical analysis[38]. The result shows that DeepRetention has significantly more interactions with known AD-associated genes than randomly selected genes ($p$-value $< 1.0 \times 10^{-6}$). The significant test results is shown in Fig. 5a. In ADBrainNexus, we calculate the number of interactions between the parent genes of differential IRs and known AD-associated genes, and calculate the number of interactions between randomly selected $k$ genes and known AD-associated genes, denoted by $t_{observed}$ and $t_{random}$, respectively. By repeating the random sampling, we obtain $10^6$ values of $t_{random}$. We then calculate $p$-value $= N_{sig}/10^6$




(a) (b)

**Fig. 5 The differential IR events predicted by DeepRetention are evaluated for biological significance. (a) The parent genes of differential IRs on the BBUKY dataset show significant functional associations with known AD-associated genes in the ADBrainNexus network (The $p$-values for both interaction genes and edges are less than $1.0 \times 10^{-6}$). In the figure, the red line represents the number of interactions between the parent genes of differential IRs and known AD-associated genes. The yellow bar chart represents the number of interaction pairs between randomly selected genes and known AD-associated genes. (b) The result of the GO enrichment on the BBUKY dataset. The parent genes of differential IRs predicted by DeepRetention are used as the input for GO enrichment in the biological processes. The enrichment analysis is performed by Metascape using the default parameters.**

where $N_{sig}$ is the number of times that $t_{random}$ value is larger than $t_{observed}$.

The functions carried out by the retained introns can be revealed by enrichment analysis[39]. We evaluate the parent genes of differential IRs on the BBUKY dataset using functional enrichment analysis. The enrichment analysis in the biological processes is performed by Metascape[40] using the default parameters. By enriching the parent genes of differential IRs (Fig. 5b), Gene Ontology (GO)[41] terms are related to AD functions, such as glial cell differentiation (GO:0010001)[42–44], regulation of neuronal synaptic plasticity (GO:0048168)[45–47], and protein phosphorylation (GO:0006468)[48]. The previous researches report that the AD mechanism of molecular control is related to glial cell differentiation[42,43]. The sustained rescue of cortical neurons by glial cell differentiation suggests that glial cell-based repair plays a beneficial role in AD[44]. The pathogenesis of AD is related to regulation of neuronal synaptic plasticity[45,46], and the improvement of cognition function of AD mice may be linked to an up-regulation of neuronal synaptic plasticity[47]. Adusumalli et al.[48] revealed that by affecting pathways involved in protein homeostasis (e.g., protein phosphorylation), changes of IR pattern may regulate the transition of physiological state from health to AD.

In summary, the differential IRs are found to be significantly associated with AD, and the parent genes of differential IRs are enriched in AD-related functions. Therefore, these results imply that DeepRetention is capable of uncovering biologically meaningful IR events.

## 4 Conclusion

In this paper, we propose DeepRetention, which detects IR by modeling the pattern of sequencing depth profiles of intronic regions. As illustrated in this work, the incorporation of distribution patterns can improve the accuracy of IR detection. IRFinder and iREAD cannot detect IR from the third-generation sequencing dataset, while DeepRetention achieves good performance when applied to third-generation long-read RNA-seq data. We explore the characteristics of retained introns predicted by DeepRetention. Compared with non-retained introns, the retained introns have significantly weaker splicing strength and higher GC content, which is consistent with the previous reports[7,12,26]. In addition, we find that the RBP-binding score of retained introns is significantly lower than that of non-retained introns, which is not uncovered by previous studies. Further, we show that DeepRetention is capable of uncovering biologically meaningful IR events by applying it to detect IR on an AD dataset. The differential IRs are functionally related to AD in an AD-specific brain functional gene network. The parent genes of differential IRs are enriched in AD-related functions. In the future, we intend to generate a gold standard dataset of IR to eliminate the limitation that the training data of DeepRetention depends on other IR detection methods. Moreover, the model structure of DeepRetention will be updated. For example, the use of attention mechanisms on the convolution module[49] or the use of the more advanced convolution architecture (such as Inception)[50] may be able to better capture the sequencing depth distribution pattern of IR. In summary, DeepRetention detects IR from a new angle of view which provides a valuable tool for IR research.

### Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2022ZD0213700) and the Natural Science Foundation of Changsha (No. kq2202105).

### References

[1] 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.

[2] D. Sulakhe, M. D'Souza, S. Wang, S. Balasubramanian, P. Athri, B. Q. Xie, S. Canzar, G. Agam, T. C. Gilliam, and N. Maltsev, Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources, *Brief. Bioinform.*, vol. 20, no. 5, pp. 1754–1768, 2019.

[3] H. D. Li, G. S. Omenn, and Y. F. Guan, A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling, *Brief. Bioinform.*, vol. 17, no. 6, pp. 1024–1031, 2016.

[4] Q. F. Chen, W. Li, P. H. Wu, L. J. Shen, and Z. L. Huang, Alternative splicing events are prognostic in hepatocellular carcinoma, *Aging* (*Albany NY*), vol. 11, no. 13, pp. 4720–4735, 2019.

[5] 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.

[6] M. Parra, W. G. Zhang, J. Vu, M. Dewitt, and J. G. Conboy, Antisense targeting of decoy exons can reduce intron retention and increase protein expression in human erythroblasts, *RNA*, vol. 26, no. 8, pp. 996–1005, 2020.

[7] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N.

Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, Widespread intron retention in mammals functionally tunes transcriptomes, *Genome Res*., vol. 24, no. 11, pp. 1774–1786, 2014.

[8] T. Ni, W. J. Yang, M. Han, Y. B. Zhang, T. Shen, H. B. Nie, Z. H. Zhou, Y. L. Dai, Y. Q. Yang, and P. C. N. Liu, et al., Global intron retention mediated gene regulation during $CD4^+$ T cell activation, *Nucleic Acids Res*., vol. 44, no. 14, pp. 6817–6829, 2016.

[9] D. X. Zhang, Q. Hu, X. Z. Liu, Y. B. Ji, H. P. Chao, Y. Liu, A. Tracz, J. Kirk, S. Buonamici, 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.

[10] 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.

[11] A. C. Smart, C. A. Margolis, H. Pimentel, M. X. He, D. N. Miao, D. Adeegbe, T. Fugmann, K. K. Wong, and E. M. Van Allen, Intron retention is a source of neoepitopes in cancer, *Nat. Biotechnol*., vol. 36, no. 11, pp. 1056–1058, 2018.

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

[13] W. Jiang and L. Chen, Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing, *Comput. Struct. Biotechnol. J*., vol. 19, pp. 183–195, 2021.

[14] 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.

[15] 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.

[16] R. Middleton, D. D. Gao, A. Thomas, B. Singh, A. Au, J. J. L. Wong, A. Bomane, B. Cosson, E. Eyras, 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.

[17] L. Broseus and W. Ritchie, S-IRFindeR: stable and accurate measurement of intron retention, *BioRxiv*, doi: 10.1101/2020.06.25.164699.

[18] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. W. Cui, G. B. Schwartz, et al., Predicting splicing from primary sequence with deep learning, *Cell*, vol. 176, no. 3, pp. 535–548.e24, 2019.

[19] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. Mccarthy, R. M. Davies, et al., Twelve years of SAMtools and BCFtools, *Gigascience*, vol. 10, no. 2, p. giab008, 2021.

[20] H. Wang, Garbage recognition and classification system based on convolutional neural network vgg16, in *Proc. 2020 3rd Int. Conf. on Advanced Electronic Materials, Computers and Software Engineering* (*AEMCSE*), Shenzhen, China, 2020, pp. 252–255.

[21] A. Wiranata, S. A. Wibowo, R. Patmasari, R. Rahmania, and R. Mayasari, Investigation of padding schemes for faster R-CNN on vehicle detection, in *Proc. 2018 Int. Conf. on Control, Electronics, Renewable Energy and Communications* (*ICCEREC*), Bandung, Indonesia, 2018, pp. 208–212.

[22] Q. Le and T. Mikolov, Distributed representations of sentences and documents, in *Proc. 31st Int. Conf. on Machine Learning*, Beijing China, 2014, pp. II-1188–II-1196.

[23] 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.

[24] B. Bai, C. M. Hales, P. C. Chen, Y. Gozal, E. B. Dammer, J. J. Fritz, X. S. Wang, Q. W. Xia, D. M. Duong, C. Street, et al., U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 41, pp. 16562–16567, 2013.

[25] R. E. Workman, A. D. Tang, P. S. Tang, M. Jain, J. R. Tyson, R. Razaghi, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, et al., Nanopore native RNA sequencing of a human poly(A) transcriptome, *Nat. Methods*, vol. 16, no. 12, pp. 1297–1305, 2019.

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

[27] G. Yeo and C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, *J. Computat. Biol*., vol. 11, nos. 2&3, pp. 377–394, 2004.

[28] Z. Birnbaum, On a use of the Mann-Whitney statistic, in *Proc. 3rd Berkeley Symp. on Mathematical Statistics and Probability*, *Volume 1*: *Contributions to the Theory of Statistics*, Berkeley, CA, USA, 2020, pp. 13–17.

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

[30] J. Yao, D. Ding, X. P. Li, T. Shen, H. H. Fu, H. Zhong, G. Wei, and T. Ni, Prevalent intron retention fine-tunes gene expression and contributes to cellular senescence, *Aging Cell*, vol. 19, no. 12, p. e13276, 2020.

[31] Z. X. Lu, Q. Huang, J. W. Park, S. H. Shen, L. Lin, C. J. Tokheim, M. D. Henry, and Y. Xing, Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization, *Mol. Cancer Res*., vol. 13, no. 2, pp. 305–318, 2014.

[32] I. Paz, I. Kosti, M. Jr. Ares, M. Cline, and Y. Mandel-Gutfreund., RBPmap: A web server for mapping binding sites of RNA-binding proteins, *Nucleic Acids Res*., vol. 42,

no. W1, pp. W361–W367, 2014.

[33] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration, *Brief. Bioinform.*, vol. 14, no. 2, pp. 178–192, 2013.

[34] J. T. Robinson, H. Thorvaldsdóttir, A. M. Wenger, A. Zehir, and J. P. Mesirov, Variant review with the integrative genomics viewer, *Cancer Res.*, vol. 77, no. 21, pp. e31–e34, 2017.

[35] J. Liu, M. Li, W. Lan, F. X. Wu, Y. Pan, and J. X. Wang, Classification of Alzheimer's disease using whole brain hierarchical network, *IEEE/ACM Trans. Computat. Biol. Bioinform.*, vol. 15, no. 2, pp. 624–632, 2018.

[36] C. X. Lin, H. D. Li, C. Deng, S. Erhardt, J. Wang, X. Q. Peng, and J. X. Wang, AlzCode: A platform for multiview analysis of genes related to Alzheimer's disease, *Bioinformatics*, vol. 38, no. 7, pp. 2030–2032, 2022.

[37] 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.

[38] C. X. Lin, H. D. Li, C. Deng, W. S. Liu, S. Erhardt, F. X. Wu, X. M. Zhao, Y. F. Guan, J. Wang, D. F. Wang, et al., An integrated brain-specific network identifies genes associated with neuropathologic and clinical traits of Alzheimer's disease, *Brief. Bioinform.*, vol. 23, no. 1, p. bbab522, 2022.

[39] Z. P. Wu, J. T. Zheng, and H. D. Li, Identification of disease-associated genes based on differential intron retention, in *Proc. 2020 IEEE Int. Conf. on Bioinformatics and Biomedicine* (*BIBM*), Seoul, Republic of Korea, 2020, pp. 228–231.

[40] Y. Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, *Nat. Commun.*, vol. 10, no. 1, p. 1523, 2019.

[41] The Gene Ontology Consortium, The gene ontology resource: 20 years and still GOing strong, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, 2019.

[42] R. L. Schnaar, Gangliosides of the vertebrate nervous system, *J. Mol. Biol.*, vol. 428, no. 16, pp. 3325–3336, 2016.

[43] M. Fang, P. Zhang, Y. X. Zhao, and X. Y. Liu, Bioinformatics and co-expression network analysis of differentially expressed lncRNAs and mRNAs in hippocampus of APP/PS1 transgenic mice with Alzheimer disease, *Am. J. Transl. Res.*, vol. 9, no. 3, pp. 1381–1391, 2017.

[44] D. W. Hampton, D. J. Webber, B. Bilican, M. Goedert, M. G. Spillantini, and S. Chandran, Cellmediated neuroprotection in a mouse model of human tauopathy, *J. Neurosci.*, vol. 30, no. 30, pp. 9973–9983, 2010.

[45] H. L. Lu, L. Liu, S. Han, B. B. Wang, J. Qin, K. L. Bu, Y. Z. Zhang, Z. Z. Li, L. N. Ma, J. Tian, et al., Expression of tiRNA and tRF in *APP/PS1* transgenic mice and the change of related proteins expression, *Ann. Transl. Med* ., vol. 9, no. 18, p. 1457, 2021.

[46] M. S. Unger, E. Li, L. Scharnagl, R. Poupardin, B. Altendorfer, H. Mrowetz, B. Hutter-Paier, T. M. Weiger, M. T. Heneka, J. Attems, et al., $CD8^+$ T-cells infiltrate Alzheimer's disease brains and regulate neuronal-and synapse-related gene expression in APP-PS1 transgenic mice, *Brain Behav. Immun.*, vol. 89, pp. 67–86, 2020.

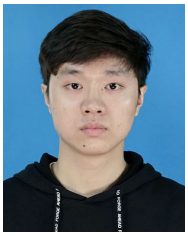
[47] Y. B. Cui, S. S. Ma, C. Y. Zhang, W. Cao, M. Liu, D. P. Li, P. J. Lv, Q. Xing, R. N. Qu, and N. Yao, Human umbilical cord mesenchymal stem cells transplantation improves cognitive function in Alzheimer's disease mice by decreasing oxidative stress and promoting hippocampal neurogenesis, *Behav. Brain Res.*, vol. 320, pp. 291–301, 2017.

[48] S. Adusumalli, Z. K. Ngian, W. Q. Lin, T. Benoukraf, and C. T. Ong, Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease, *Aging Cell*, vol. 18, no. 3, p. e12928, 2019.

[49] S. H. Wang, Q. H. Zhou, M. Yang, and Y. D. Zhang, ADVIAN: Alzheimer's disease VGG-inspired attention network based on convolutional block attention module and multiple way data augmentation, *Front. Aging Neurosci.*, vol. 13, p. 687456, 2021.

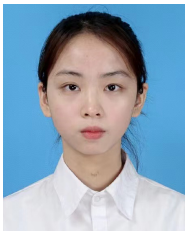
[50] Y. Yan, X. J. Yao, S. H. Wang, and Y. D. Zhang, A survey of computer-aided tumor diagnosis based on convolutional neural network, *Biology* (*Basel*), vol. 10, no. 11, p. 1084, 2021.

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



**Jiantao Zheng** received the BEng and MEng degrees from Zhengzhou University and Central South University, China in 2018 and 2021, respectively. His research area is intron retention detection.

**Hong-Dong Li** received the BEng (in pharmaceutical engineering) and PhD (in analytical chemistry) degrees from Central South University, 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), and *Chemom. Intell. Lab. Syst.* According to Google Scholar, his publications have received over 3800 citations with H-index=29. The maximum citation number of a single paper is over 660. 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 interest includes bioinformatics, chemometrics, and machine learning.



**Jiashu Liu** received the BEng degree from Central South University, China in 2021. She is currently a master student in bioinformatics at Central South University. Her research area is intron retention detection.

**Cuixiang Lin** received the BEng and MEng degrees both from Central South University, China in 2008 and 2010, respectively. She is a PhD candidate at Central South University. Her research interests include identification of biomarkers of disease and disease mechanism.