Research Article

Investigating Alzheimer's Disease-Associated Genes Using Differential Splicing Frequency Analysis

Yang Yao[1], Zhi Cheng[1], Shunmei Chen[2], Jingsong Shi[3], Yiyao Zhang[1], Chang Liu[4], Dongsheng Wei[1], Tao Zhang[1], Guangyou Duan[5], Shan Gao[1]

1. College of Life Sciences, Nankai University, China; 2. Biomedical Engineering Research Institute, Kunming Medical University, China; 3. National Clinical Research Center for Kidney Diseases, Nanjing University, China; 4. School of Medicine, Nankai University, China; 5. School of Life Sciences, Qilu Normal University, Jinan, China

Accurately quantifying the expression of individual isoforms remains a formidable challenge, especially in contexts like neurodegenerative diseases and cancers, which are noted for their high isoform diversity. The first contribution of the present study is the development of a new method, named differential splicing frequency analysis (DSFA), which enables more sensitive capture of the alternative splicing information in RNA-seq data. DSFA quantifies the expression levels of splicing junctions, rather than those of genes, isoforms or exons. Application of DSFA to the analysis of Alzheimer's disease (AD)-associated genes resulted in the identification of APP/58417N and APP/52804N as differentially expressed splicing junctions in human and mouse, respectively. Observed in a considerable portion of the re-analyzed RNA-seq datasets, the splicing frequencies of APP/58417N and APP/52804N were significantly decreased in AD groups. Such reductions in splicing frequencies lead to decreased production of secretory APP proteins, potentially playing a critical role in AD onset or progression. The present study has proposed over-expression of U1 snRNA as an effective method for modulating the splicing frequencies of splicing junctions, thereby rapidly establishing cellular or animal AD models. Therefore, the present study has provided new methods, preliminary results, and valuable insights, advancing the understanding of the functions of U1 snRNA and the roles of alternative splicing in its associated diseases.

Yang Yao and Zhi Cheng equally contributed to this work.

 doi.org/10.32388/CNV1ZB

Corresponding authors: Shan Gao, gao_shan@mail.nankai.edu.cn; Guangyou Duan, guangyou.duan@qlnu.edu.cn

Introduction

RNA sequencing (RNA-seq), a high-throughput gene expression measurement technology, is widely utilized in life sciences due to its ability to provide a comprehensive understanding of the transcriptome. The transcriptome is traditionally defined as encompassing mRNAs and long non-coding RNAs (lncRNAs) exceeding 200 nucleotides in length. RNA-seq outperforms microarray in terms of specificity, thereby enhancing the ability to discriminate between various isoform[1]. However, many researchers still do not distinguish between different isoforms of a gene when quantifying gene expression using RNA-seq data, particularly in differential gene expression analysis (DGEA). This may inadvertently neglect the potential significant differences in isoform expression, when the total expression level of the gene is not significantly different. On the other hand, other researchers employ advanced bioinformatics tools to quantify the expression of individual isoforms. These tools rely on predictive models rather than direct quantification, due to the limited read lengths characteristic of RNA-seq data[2]. The predictive models are founded on algorithms to allocate reads to the most probable isoform, taking into account various factors, including read coverage and known transcript structures. A previous study[3] concluded that none of the evaluated methods is accurate enough for *de novo* transcriptome assembly, where isoform structures need to be inferred directly from the RNA-Seq data. Consequently, accurately quantifying the expression of individual isoforms remains a formidable challenge, especially in contexts like neurodegenerative diseases[4] and cancers[5], which are noted for their high isoform diversity. For example, a recent study concentrating on tau pathology[6] reported an unexpectedly high degree of isoform diversity and alternative splicing events across various genes. Notably, a total of 1408 isoforms were identified from the gene *APOE*, with only 13 being previously characterized, while the remaining 1395 were newly discovered.

Alternative technologies indeed offer improved differentiaion of individual isoforms, addressing the challenges posed by the high isoform diversity. Tiling array and HIDE-seq[7], as such technologies, have been used for a comprehensive view of the genome at the isoform resolution. The more promising technologies are long-read sequencing technologies, including Pacific Biosciences′ full-length transcriptome sequencing (PacBio cDNA-seq) and Oxford Nanopore Technologies′ direct RNA sequencing (Nanopore RNA-seq)[8]. These long-read sequencing technologies generate long reads, which, in theory,

allow for the direct quantification of nearly all isoforms of a gene. However, they still have limitations, including lower throughput, lower sequence fidelity, or higher costs[9]. As a result, the majority of RNA-Seq studies continue to rely on short-read sequencing, a trend that is likely to persist until competing technologies mature. Rather than adopting alternative technologies, there is potential in developing new computational methods designed to fully exploit the alternative splicing (AS) information present in RNA-seq data.

To more sensitively capture the AS information in RNA-seq data, we developed a new method that quantifies the expression levels of splicing junctions, rather than those of genes, isoforms or exons. This method identifies splicing junctions with significant differences in their expression levels between two or more groups or conditions, which are termed as differentially expressed (DE) splicing junctions (**Table 1**). As DE splicing junctions are analogous to DE genes identified by the conventional DGEA, the new method was named differential splicing frequency analysis (DSFA). DGEA focuses specifically on the identification of DE genes, whereas DSFA focuses specifically on the identification of DE splicing junctions within a gene. We then applied DSFA in our AD research. This application served not only to exemplify the utility of DSFA, but also to identify DE splicing junctions between AD groups and their respective controls or wild types. As part of our AD research, the present study began with the application of DGEA to AD-associated genes using publicly available RNA-seq data, followed by validation of the primary findings using neurons differentiated from human embryonic stem cells (hESCs).

Results

Differential splicing frequency analysis

To facilitating the development of a new method, we designed Nankai naming system for naming genomic features. Nankai naming system assigns names to genomic features based on their length, using the format Gene/Ex for exons, Gene/Iy for introns and Gene/yN for splicing junctions, where x and y represent their respective lengths in base pair (bp). The new system ensures that each exon and intron can be distinctly identified, with the potential for extremely rare overlaps in naming. For example (**Figure 1**), APP/I58417 denotes a unique intron of the gene *APP*, which spans 58,417 bp in human genome. This intron is typically present as the first intron in certain long *APP* transcripts (**Detailed below**), defined as those longer than half the length of the longest APP transcript. Correspondingly, APP/58417N denotes a unique splicing junction, where the two flanking exons of APP/I58417 are joined in the mature RNA, with

APP/I58417 having been excised. Splicing junctions can be identified by counting splicing-junction (sj) reads that align properly to these junctions, adhering to a set of criteria (**Materials and methods**) designed to minimize false positives. Thus, a matrix composed of sj read counts (termed as jrc matrix) can be derived from an RNA-seq dataset to represent the expression levels of splicing junctions in the examined samples. As the sj read count directly correlates with the number of splicing events at a splicing junction, each column in a jrc matrix represents the numbers of splicing events at all identified splicing junctions within an individual sample, thereby characterizating the sample by a set of occurrences of splicing junctions in all or selected genes, named as a splicing occurrency profile or signature, respectively; and each row represents the numbers of splicing events at a specific splicing junction across all examined samples.

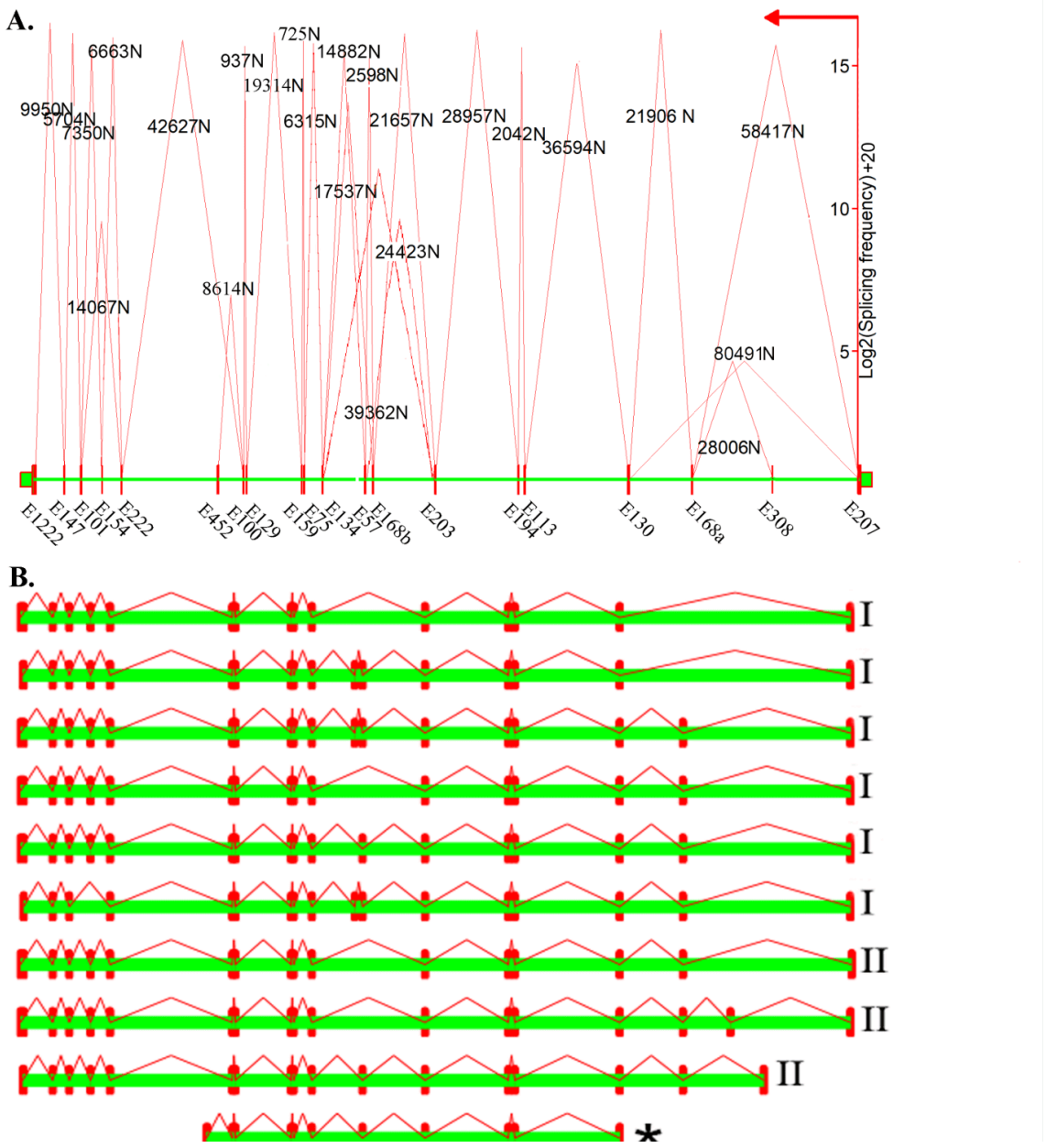

Figure 1. Understanding splicing frequency and its analysis

The exons and introns are indicated in red and green, respectively. A splicing junction is designated using the name format Gene/yN, where y is the size of its corresponding intron. APP/I58417 is typically present as the first intron in certain long *APP* transcripts. (**A**) The RNA-seq data of a sample containing hESC-derived neurons has been deposited in the NCBI SRA database under the accession number SRR8433687. Using this data, 20 exons and 24 known splicing junctions of the *APP* gene (Table 1) are represented in a left-to-right orientation. The normalized expression level of each splicing junction is represented by the

doi.org/10.32388/CNV1ZB

height of the zigzag line (in red color) above the intron region, which can also be interpreted as the relative frequency of splicing events, termed as splicing frequency. The y-axis represents the log value of splicing frequency plus 20. Among the 24 splicing junctions, five (APP/80491N, APP/28006N, APP/24423N, APP/8614N, and APP/14067N) are considered rare, as their values fall below 10 on the y-axis. (**B**) Nine long *APP* transcripts can be classified into type 1 and type 2 (indicated by I and II) ones. One transcript (Ensembl: ENST00000448850) is annotated as being transcribed from a transcription initiation site (indicated by *) downstream of APP/I58417.

In a normalized jrc matrix (**Materials and methods**), the scaled expression level of a splicing junction can also be interpreted as the relative frequency of splicing events that have occurred at this splicing junction, hereinafter referred to as splicing frequency. Then, DSFA (**Introduction**) can be used to identify splicing junctions with significant different splicing frequencies, *i.e.*, DE splicing junctions (**Table 1**). DSFA is primarily applied to non-DE genes, which are usually neglected in further analyses following the application of DGEA. DSFA can also be applied to DE genes, but the identified non-DE splicing junctions of DE genes should be interpreted with caution, as they contribute to the DE genes. When DSFA is applied to non-DE genes, the comparison of splicing frequencies of splicing junctions should be performed exclusively within the same genes. This approach ensures that the analysis is focused on the AS events specific to each gene. In contrast, when DSFA is applied to DE genes, the comparison of splicing frequencies of splicing junctions across genes is theoretically possible, although it remains an open question. However, many non-DE genes identified using one dataset could be identified as DE genes between two same conditions using another dataset, and vice versa.

Splicing junctions	logFC	logCPM	PValue	FDR
APP/69N*	8.009573	8.765664	7.34E-16	**1.98E-14**
APP/725N	0.025445	15.81392	0.426282	0.595971
APP/734N	0.508635	6.917798	0.781577	0.811637
APP/937N	0.057404	15.64494	0.069576	0.234819
APP/2042N	0.073655	15.63724	0.046497	0.234819
APP/2598N	0.066097	15.35966	0.214842	0.446386
APP/5198N*	6.671129	7.640209	2.58E-07	**3.49E-06**
APP/5704N	-0.03109	16.04939	0.357988	0.571415
APP/6315N	-0.02098	15.67411	0.515372	0.632502
APP/6663N	-0.03989	15.6045	0.324961	0.571415
APP/7350N	-0.03392	15.5967	0.403781	0.595971
APP/8614N	0.493416	7.350774	0.664004	0.747004
APP/9950N	-0.05928	16.37756	0.06267	0.234819
APP/14067N	0.827405	10.18114	0.00805	0.054339
APP/14882N	0.046963	15.44073	0.35978	0.571415
APP/17537N	0.226791	13.80825	0.311978	0.571415
APP/19314N	0.040544	16.11583	0.180124	0.442123
APP/21657N	0.002798	16.0602	0.927489	0.927489
APP/21906N	-0.05933	16.15769	0.097023	0.291068
APP/24423N	-0.93972	8.239727	0.068857	0.234819
APP/28957N	0.040463	16.2244	0.214927	0.446386
APP/36594N	0.020258	15.14655	0.753715	0.811637
APP/39362N	0.198866	11.02213	0.477235	0.613588
APP/42627N	0.023854	15.8063	0.44146	0.595971
APP/56921N*	4.107257	6.180819	0.129143	0.348686

doi.org/10.32388/CNV1ZB

Splicing junctions	logFC	logCPM	PValue	FDR
APP/58417N	-0.22646	15.57042	3.95E-05	**0.000356**
APP/80491N	-0.72722	6.843774	0.56086	0.658401

Table 1. Splicing junctions of the *APP* gene

*Analogous to differential gene expression analysis (**Introduction**), differential splicing frequency analysis (DSFA) was designed to identify significantly differential expressed (DE) splicing junctions. In the present study, the expression of 27 splicing junctions was compared between cell groups over-expressing U1 snRNA and their controls. Among these 27 splicing junctions, 24 known splicing junctions were frequently detected across different samples, while the remaining three novel splicing junctions (indicated by *) are rare splicing junctions. APP/69N and APP/5198N were identified as DE splicing junctions by DSFA, along with APP/58417N. LogFC (Log Fold Change) is the logarithmic (base 2) transformation of the fold change between two conditions or groups; LogCPM (Log Counts Per Million) is the logarithm of the normalized gene expression values, typically calculated as counts per million (CPM); PValue (P-value) and FDR (False Discovery Rate), also referred to as the adjusted P-value, were calculated using the edgeR package (**Materials and methods**).*

The first splicing junction of APP identified as differentially expressed

Upon re-examining the previous results derived from RNA-seq datasets obtained from NCBI SRA database (**Materials and methods**), it is evident that the previous studies using DGEA failed to identify a consistent pattern of up- or down-regulation for AD-associated genes in AD groups across these datasets. Notably, six AD-associated genes (*APP*, *PSEN1*, *PSEN2*, *APOE*, *CLU*, and *MAPT*) were often identified as non-DE genes in previous studies using DGEA. In the present study, we applied DSFA to the six AD-associated genes by re-analyzing the selected RNA-seq datasets (**Supplementary 1**). In these RNA-seq datasets, the AD groups primarily consist of tissues from AD patients or model mice (*e.g.*, 5xFAD). Overall, our analysis revealed several findings that were inconsistent with those reported in the previous study on Tau pathology[6]. First, *APP* was identified as having the largest number of splicing junctions (**Table 1**) among the six AD-associated genes in the present study. Furthermore, *APP* accounted for nearly all of the identified novel and DE splicing junctions. In contrast, the previoust study[6] reported that the largest number of isoform types was transcribed from *APOE*, rather than *APP*. Second, although

both the present study and the previous study[6] identified a substantial number of novel splicing junctions within the AD-associated genes, the present study revealed that almost all of these novel splicing junctions were supported by a limited number of sj reads. These junctions, termed as rare splicing junctions, have unclear research value. Three such rare splicing junctions, APP/56921N, APP/5198N, and APP/69N (**Table 1**), occurred indiscriminately in AD and wild-type or control groups with very low frequency and were analyzed in detail below. Therefore, the inconsistency between the present study and the previous study[6] lies in the fact that the previous study incorporated a much larger number of rare splicing junctions into the analysis.

SRA ACC	Species	Tissue/Cell	Samples (Number)	DE splicing junction
SRP543852	Mouse	Hippocampus	AD(3), W/C(3)	APP/52804N
SRP447810	Mouse	Hippocampus	AD(6), W/C(6)	APP/52804N
SRP555827	Mouse	Hippocampus	AD(3), W/C(3)	NA
SRP521327	Mouse	Hippocampus	AD(3), W/C(3)	NA
SRP490059	Human	iPSC-derived neurons	AD(28), W/C(24)	APP/58417N
SRP382946	Human	iPSC-derived neurons	AD(3), W/C(9)	APP/58417N
SRP517721	Human	Microglia	APOE-KO(5), APOE2(5), APOE3(4), APOE4(5)	APP/58417N
ERP161086	Human	Cerebellum and prefrontal cortex	AD(13), D&AD(14), W/C(15)	APP/58417N*
SRP540706	Human	Hippocampus, SN, CG, PC, and TC	AD(30), W/C(60), LBD(30)	APP/58417N
SRP178463	Human	hESC-derived neurons	U1 over-expressed(2), W/C(2)	APP/58417N

Table 2. Selected RNA- seq datasets for re-analysis

The RNA-seq data presented in this table were derived from a total of 199 human and 194 mouse datasets. Detailed information regarding the remaining datasets and the corresponding results are provided in

Supplementary 1 and 2, respectively. The dataset (SRA: RP1784630) generated from experiments in the present study was used to validate the primary finding, while other data generated from previous studies was re-analyzed for data mining purposes, which collectively contributed to the identification and validation of the primary finding. SRA ACC: accession number in NCBI SRA database; DE splicing junction: differentially expressed splicing junctions between AD and control or wild-type samples; iPSC: Induced Pluripotent Stem Cell; hESC: human embryonic stem cell; SN:Subtantia nigra; CG:Cingulate gyrus; PC:Parietal cortex; TC:Temporal cortex; W/C: wild type for variants or controls for patients; AD: Alzheimer's disease; D&AD: Down syndrome with Alzheimer's disease; LBD: Lewy body dementia; APOE-KO: human microglials with APOE knockout; APOE2-4: human microglials expressing the type 2-4 APOE gene; U1 over-expressed: hESC-derived neurons over-expressing U1 snRNA.

According to the annotations of human genome (**Materials and methods**), two distinct classes of long APP transcripts, have been identified and designated as Class 1, and 2 in the present study (**Figure 1**). These classes are characterized by their first exons and first splicing junctions that are located downstream of the first exons. The first splicing junctions of Class 1 long transcripts are APP/80491N or APP/58417N, which are formed through the excision of APP/I80491 and APP/I58417 — the largest and second-largest introns among all introns in the six AD-associated genes, respectively, while those of Class 2 long transcripts are APP/58838N, APP/41847N, or APP/28006N. The first exons of Class 1 long transcripts encode a signial peptide MLPGLALLLLAAWTAR|ALE (the vertical line indicates the cleavage break point) that is responsible for the secretion of APP proteins, whereas those of Class 2 long transcripts do not encode a signial peptide. The first exons of Class 2 long transcripts containing APP/28006N encode MDQLEDLLVLFINY, which was identified as a non-signal peptide, while those of Class 2 long transcripts containing APP/58838N or APP/41847N do not encode any peptide. In contrast to human, only Class 1 long *APP* transcripts have been identified in mouse, which is characterized by its unique splicing junction APP/52804N. The first exons of these mouse *APP* transcripts encode a signial peptide MLPSLALLLLAAWTVR|ALE. Expanded analysis revealed that the signal peptides of APP homologs are highly conserved among animals, despite considerable divergence in their amino acid (aa) sequences. A notable example is the signial peptide of APP (UniProt: P14599) in *Drosophila melanogaster* (fruit fly) MCAALRRNLLLRSLWVVLAIGTAQVQ|A. Furthermore, Class 1 long transcripts containing APP/58417N are predominantly responsible for the translation of secretory APP proteins, while other Class 1 long transcripts, Class 2 long transcripts, and short transcripts are unlikely to make a significant

contribution due to their exceedingly low expression levels. Therefore, the first exons encoding a signal peptide could play a crucial role in APP functions.

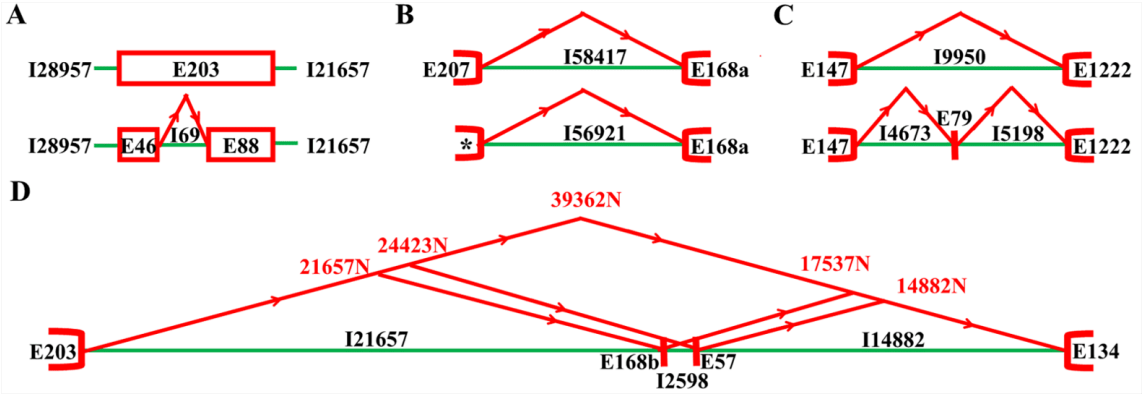

Figure 2. Alternative splicing induced by over-expression of U1 snRNA

The exons and introns are indicated in red and green, respectively. Over-expression of U1 snRNA increases the splicing frequencies of rare splicing junctions APP/56921N, APP/5198N, and APP/69N. APP/58417N, APP/9950N, APP/21657N, APP/2598N, and APP/14882N are high-frequency splicing junctions. APP/E168a, and APP/168b are two different exons (Figure 1A). When the smallers and rare introns are formed in RNA splicing, (**A**) APP/I58417 is split into APP/I56921 and a novel exon (indicated by *), which encodes a new peptide MLRSCLHDSWARGGCISLRTS; (**B**) APP/I9950 is split into APP/I4673, APP/E79, and APP/I5198; and (**C**) APP/E203 is split into APP/E46, APP/I69, and APP/E88. When APP/I58417, APP/I9950, and APP/E203 are formed, their splicing types are alternative 5′ splice site, exon skipping and intron retention, respectively. (**D**) The transcript encoding APP695 contains APP/39362N; the transcript encoding APP751 contains APP/21657N and APP/17537N; and the transcript encoding APP770 contains APP/21657N, APP/2598N and APP/14882N.

The primary finding of the present study was the identification of APP/58417N and APP/52804N as DE splicing junctions in human and mouse, respectively. Specifically, the splicing frequencies of APP/58417N and APP/52804N were significantly decreased in AD groups, which was consistently observed in a considerable portion of the re-analyzed RNA-seq datasets derived using hippocampus tissue or differentiated neurons. As APP/58417N are APP/52804N are the first splicing junctions of Class 1 long *APP* transcripts, this reduction indicated a decreased inclusion of the first exons encoding a signal peptide (**Figure 2**). Besides encoding a signal peptide, these first exons (*e.g.*, APP/E207 for human) exhibit another

notable feature: an exceptionally high enrichment of CGG repeats compared to other regions of the *APP* gene. Expanded CGG-repeats have been linked to neurodevelopmental and neurodegenerative disorders[10]. However, no ac4C modification was detected on these CGG repeats by re-analysis of a RIP-seq dataset (SRA: SRP544711), despite their alignment with the CXX motif. In conclusion, the reduced splicing frequencies of APP/58417N and APP/52804N lead to decreased production of the secretory forms of APP proteins, potentially playing a critical role in AD onset or progression.

Further analysis with experimental validation

The decrease in the splicing frequencies of the first splicing junctions in *APP* could be attributed to multiple factors, including alternative 5′ splice site, alternative transcription initiation sites (TISs) within or downstream of APP/I58417, premature termination of transcription, or splicing defects between its adjacent exons. According to the annotations of human genome (**Materials and methods**), a subset of short APP transcripts (*e.g.*, Ensembl: ENST00000448850) are transcribed from TISs downstream of APP/I58417 (**Figure 1**). Additionally, several APP transcripts contain alternative 5′ splice sites, such as those upstream of APP/28006N or APP/56921N (**detailed below**). However, the exceedingly low expression levels of these transcripts are insufficient to account for the significant decrease of APP/58417N. One potential contributing factor could be premature cleavage and polyadenylation (PCPA) [11]. However, PCPA was not detected in the re-analyzed RNA-seq datasets, likely due to the limitations of technologies, which may not be sensitive enough to capture such events. To address this, we turned to publicly available PacBio cDNA-seq data (*e.g.*, SRA: SRP067402) and HIDE-seq[7], due to their higher sensitivities and also used PARE-seq (SRA: SRP007508), CAGE-seq (SRA: DRP000372), GRO-seq (SRA: SRP095085), and PA-seq (SRA: SRP017395) data in the analysis. However, no solid evidence was identified to guide further analysis or experiments. Fortunately, our previous study revealed over-expression of U1 snRNA result in high similar AD-like symptoms by influencing RNA splicing. Based on this finding, we decided to investigate the mechanisms underlying the decrease in the splicing frequencies of the first splicing junctions by over-expression of U1 snRNA in cells.

Comparisions between AD groups and their controls at the tissue level generally result in only subtle expression changes of the studied genes or their splicing junctions. This is primarily due to the fact that these tissues are composed of a variety of cell types, which exhibit substantial differences in the expression levels of the same gene. For example, one of our previous studies revealed that the expression levels of *APP* and *MAPT* in oligodendrocytes were exceptionally higher than those in the other 12 cell

types, exceeding twofold the levels in neurons. Therefore, the experiments utilizing specific cell types—such as neurons—can lead to more accurate quantification of splicing frequencies for comparison. Using hESC-derived neurons (**Materials and methods**), we detected the splicing frequency of APP/58417N to be significantly decreased (logFC =-0.23, P << 0.01) in U1 snRNA over-expressed groups, similar to the effects observed in AD groups. Then, we proposed over-expression of U1 snRNA as a novel method for modulating the splicing frequencies of splicing junctions, thereby potentially generating AD models.

The experiments using hESC-derived neurons revealed that over-expression of U1 snRNA increases the splicing frequencies of rare splicing junctions APP/56921N, APP/5198N, and APP/69N. The corresponding rare introns APP/I56921, APP/I5198N, and APP/I69N are located within the relatively larger and high-frequency introns APP/I58417 and APP/I9950, as well as the exon APP/E203, respectively. When these smallers and rare introns are formed in RNA splicing, APP/I58417 is split into APP/56921N and a novel exon encoding a new peptide MLRSCLHDSWARGGCISLRTS (**Figure 2A**), which was identified as a non-signal peptide. In addition, APP/I9950 is split into APP/I4673, APP/E79, and APP/I5198 (**Figure 2B**), while APP/E203 is split into APP/E46, APP/I69, and APP/E88 (**Figure 2C**). The above findings suggested that over-expression of U1 snRNA increases the splicing frequencies of rare splicing junctions in RNA splicing, compared to their alternative high-frequency splicing junctions. This hypothesis may also provide a plausible explanation for a finding reported in a previous study[12], which compared the levels and alternative splicing of three APP transcripts in brain tissue of hAPP transgenic and nontransgenic mice and of humans with and without AD. The three APP transcripts encoding APP695, APP751, and APP770 are characterized by the presence of specific splicing junctions: APP/39362N, APP/17537N, and APP/14882N, respectively (**Figure 2D**). Specifically, frontal cortex of humans with AD showed a subtle increase in the relative abundance of APP751, compared with their controls. According to our findings, the splicing frequencies of rare splicing junctions APP/39362N and APP/17537N are increased in AD samples, compared to their alternative high-frequency splicing junctions APP/21657N and APP/14882N (**Figure 1A**), predisposing the formation of the transcripts encoding APP695 and APP751.

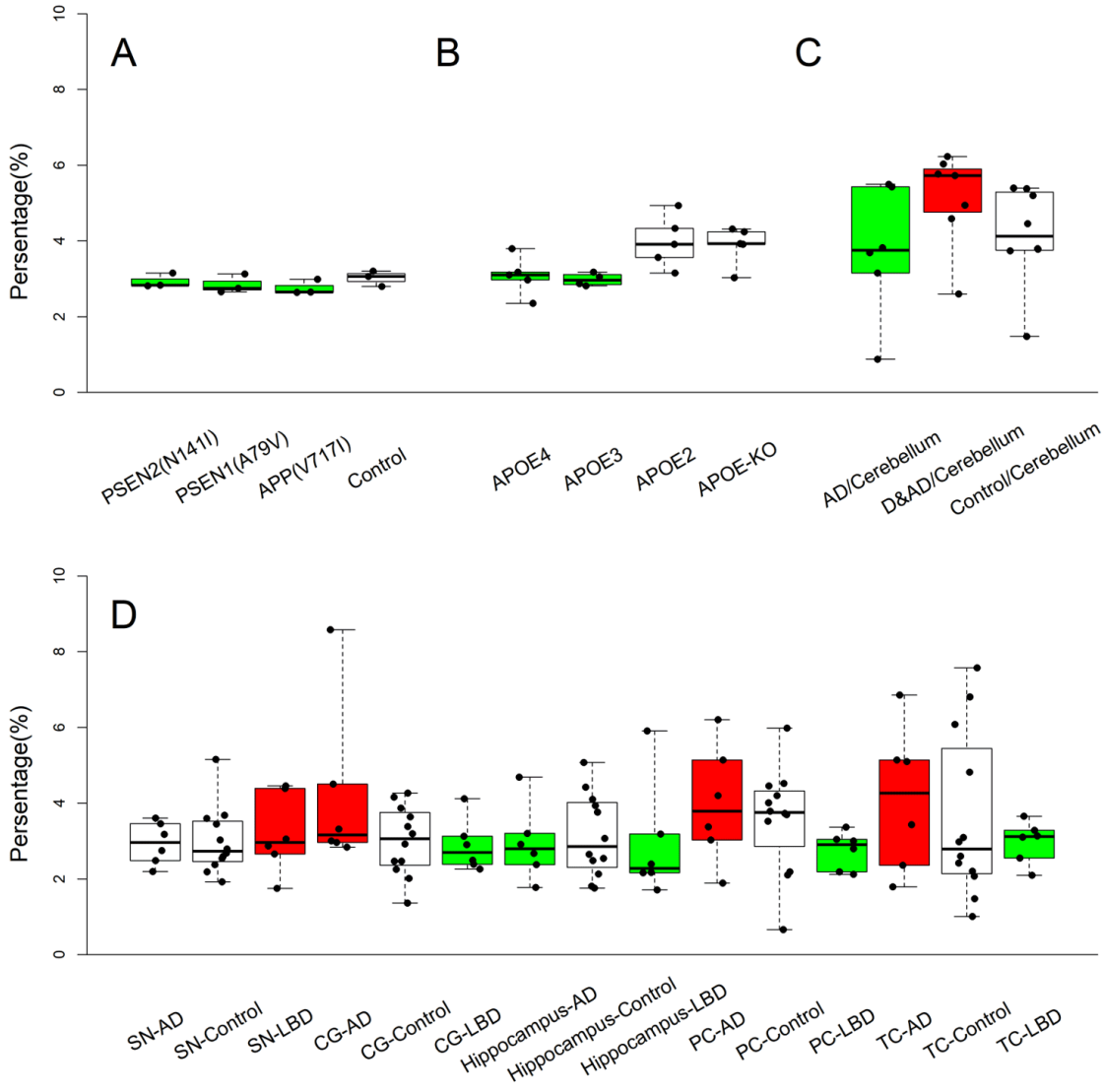

Figure 3. Expression changes of APP/58417N and APP/52804N in various tissue or cell types

The increased and decreased changes are indicated by the red and green colors of the boxes, respectively. Blank boxes represent control or no change; (**A**) Alzheimer's disease (FAD) patients with APP(V717I), PSEN1(A79V), and PSEN2(N141I) mutations. (**B**) The APOE-KO group used as the control group; (C) Cerebellum obtained from AD, D&AD patients or their controls; (**D**) Hippocampus, SN, CG, PC, and TC obtained from AD, LBD patients or their controls. Cingulate gyrus; PC:Parietal cortex; TC:Temporal cortex. W/C: wild type for variants or controls for patients; AD: Alzheimer's disease; D&AD: Down syndrome with Alzheimer's disease; LBD: Lewy body dementia; APOE-KO: human microglials with APOE knockout; APOE2-4: human microglials expressing the type 2-4 APOE gene

doi.org/10.32388/CNV1ZB

Expanded analysis of the primary finding

Expanded analysis was conducted to investigate the changes in splicing frequencies of APP/58417N and APP/52804N across various tissue or cell types under different physiological or pathological conditions using the selected RNA-seq datasets (**Supplementary 1**). By re-analysis of an RNA-seq dataset (SRA: SRP490059), we found that the splicing frequency of APP/58417N was increased during the differentiation of induced pluripotent stem cells (iPSCs) into neurons (**Table 2**). Specifically, the average splicing frequency escalated from 5.6% at 1 day post-induction (dpi) to 6.9% at 10 dpi, 7.2% at 30 dpi, and ultimately to 9% at 60 dpi. This increase in splicing frequency was accompanied by a simultaneous increase in the expression of the *APP* gene. Such a dual increase in *APP* expression and splicing frequency of APP/58417N ensures sufficient production of secretory APP proteins, which are critical for neuron differentiation. However, the splicing frequencies of APP/58417N in AD groups were expected to exhibit significant changes but did not, compared to their respective controls at 1, 10, 30, and 60 dpi. This unexpected finding may be attributed to several factors, including differences in cell type composition or tissue origin. Notably, the iPSCs used in the previous study were derived from dermal fibroblasts of patients with late-onset sporadic AD and their age-matched controls. In contrast, when iPSC were derived from familial Alzheimer's disease (FAD) patients with APP(V717I), PSEN1(A79V), and PSEN2(N141I) mutations (SRA: SRP382946), the splicing frequency of APP/58417N was detected to be significantly decreased in the neurons derived from these iPSC (**Figure 3A**), compared to the controls. In another study, the effects of the human APOE isoforms (APOE2, APOE3, APOE4) and the APOE knockout (APOE-KO) on the microglial response to amyloid beta pathology in brains of AppNL-G-F mice were investigated using RNA-seq and ATAC-seq (SRA: SRP517721). By re-analysis of the RNA-seq dataset, we detected the splicing frequency of APP/58417N to be significantly decreased in both APOE3 and APOE4 groups, relative to the APOE-KO group (**Figure 3B**). In contrast, no such significant decrease was observed in the APOE2 group.

The changes in splicing frequencies of APP/58417N and APP/52804N can serve as indicators to validate cellular or animal AD models (**Table 2**). By re-analysis of the RNA-seq datasets (SRA: SRP543852 and SRP447810), we detected the splicing frequency of APP/52804N to be significantly decreased in hippocampus from 3- and 5-month 5xFAD mice, respectively. However, no significant change was observed in hippocampus from 6- and 7-month-old 5xFAD mice using other RNA-seq datasets (SRA: SRP555827 and SRP521327), respectively. This finding suggested that establishing criteria for validating AD models is essential to avoid false positives. Investigating the expression changes of APP/58417N and

APP/52804N across various tissue or cell types provided new direction for future research. By re-analysis of an RNA-seq dataset (SRA: ERP161086), we detected the splicing frequency of APP/58417N to be significantly decreased in the cerebellum, but not in the prefrontal cortex of AD patients. In contrast, this splicing frequency was significantly increased in the cerebellum, but not in the prefrontal cortex of AD patients who also exhibited Down syndrome (**Figure 3C**). Another study (SRA: SRP540607) revealed that the tissue specificity of the splicing frequency of APP/58417N was ranked from highest to lowest as follows: temporal cortex (TC), parietal cortex (PC), subtantia nigra (SN), cingulate gyrus (CG), and Hippocampus (**Table 2**). However, the changes in splicing frequencies of APP/58417N between AD patients, Lewy body dementia (LBD) patients, and their respective controls are highly complex (**Figure 3D**). Specifically, the splicing frequency of APP/58417N was increased in TC, PC, CG of AD patients, whereas it was decreased in TC, PC, CG of LBD patients, compared to their respective controls. These preliminary results may be attributable to differences in cell type composition. Therefore, they necessitate validation through experiments using specific cell types.

Conclusion

The first contribution of the present study is the development of DSFA. Using DSFA, all publicly available RNA-seq data can be re-analyzed to detect DE splicing junctions. This strategy provides a new research direction, potentially obtaining novel findings that may have been overlooked in previous studies using the conventional DGEA. Application of DSFA to the analysis of AD-associated genes revealed that the splicing frequencies of APP/58417N and APP/52804N were significantly decreased in both AD groups and groups over-expressing U1 snRNA, compared to their respective control or wild-type groups. Such reductions in splicing frequencies lead to decreased production of secretory APP proteins, potentially playing a critical role in AD onset or progression. Future research will focus on elucidating the impacts of the comparatively increased production of non-secretory APP proteins. Follow this research direction, cellular or animal experiments will be conducted to examine the effects of over-expressing APP proteins that lack a signal peptide. These experiments are expected to provide insights into the molecular mechanisms underlying AD onset or progression.

The understanding of the functions of U1 snRNA in alternative splicing is still incomplete. Although it is already accepted that U1 snRNA interactions with RNA splice sites can modulate splicing outcomes[13], the specific mechanisms and regulatory elements involved are still under investigation. The present study has proposed over-expression of U1 snRNA as an effective method for modulating the splicing

frequencies of splicing junctions, thereby rapidly establishing cellular or animal AD models. In addition, the changes in splicing frequencies of APP/58417N and APP/52804N can serve as indicators to validate AD models. Particularly, our findings suggest that over-expression of U1 snRNA increases the splicing frequencies of rare splicing junctions in RNA splicing, compared to their alternative high-frequency splicing junctions. However, more complex factors, such as RNA structures, rather than merely the sizes of potential introns or exons, may be involved in determining the splicing frequencies of these junctions, which merits further research. Therefore, the present study has provided new methods, preliminary results, and valuable insights, advancing the understanding of the functions of U1 snRNA and the roles of alternative splicing in its associated diseases.

Materials and methods

Data acquisition and analysis

The raw HIDE-seq data were downloaded from the website (https://www.med.upenn.edu/dreyfusslab/protocols.html) for the analysis on a local server. By searching "Alzheimer's Disease" and "RNA-seq", information of 199 human and 194 mouse GEO datasets (**Supplementary 1**) was retrieved. The corresponding raw sequencing reads in FASTQ format were downloaded from the NCBI SRA databases. After the exclusion of a low-quality dataset, a total of 46 human and 46 mouse projects were selected for re-analysis using both DGEA and DSFA, based on the following criteria: (1) the exclusion of single-cell RNA-seq (scRNA-seq) and single-nucleus RNA-seq (snRNA-seq) datasets, as well as other non bulk RNA-seq data; (2) the exclusion of datasets associated with irrelevant research topic, *e.g.*, Cancre; and (3) the requirement that the average data size in a project be above 2 Gb. All the raw reads were cleaned using the pipeline Fastq_clean[14] v2.0. The cleaned reads were aligned to the human and mouse genome GRCh38 and GRCm39 using STAR v2.5.2b. For each project, a gene expression matrix was generated to represent the expression levels of all human or mouse genes, while a jrc matrix was generated to represent the splicing frequencies of splicing junctions in six AD associated genes. To account for variations in library size, the gene expression matrix was normalized by dividing each count by the total count of all genes for DGEA. In contrast, the jrc matrix (**Results**) was normalized by dividing each count by the total count of its respective gene for DSFA. DGEA and DSFA were performed between different groups using the R package edgeR v3.22.5. The analysis results obtained from all datasets using DSFA are provided in **Supplementary 2**. The significantly DE genes and

DE splicing junctions were selected based on the criteria: |log2foldchange| > 1 and pvalue<0.05. Statistics and plotting were conducted using the software R v4.3.2 with the Bioconductor packages[1]. The Perl or Python scripts to generate a jrc matrix for DSFA were included in Fastq_clean[12], which is highly recommended for cleaning and quality control of raw reads, as contaminations such as low-quality nucleotides or even very short residual adapter sequences can result in erroneous identifications of splicing junctions, thereby generating false-positive counts.

Validation using cell experiments

The STEMdiff™ Neural System (STEMCELL, USA) was employed to generate hESC-derived neurons. H1 hESCs, maintained in our laboratory, were induced for their differentiation into NPCs using the STEMdiff™ Neural Induction Medium (CATALOG 05835) and STEMdiff™ Neuron Differentiation Kit (CATALOG 08500). Subsequently, these NPCs were matured into neurons using STEMdiff™ Neuron Maturation Kit (CATALOG 08510). The identity of these neurons was validated by the expression of *TUJ*, as detected using qPCR. However, the markers *SOX2* and *OCT4* were also detected at moderate levels using qPCR, suggesting the presence of a certein portion of undifferentiated hESCs or NPCs.

Two groups of the hESC-derived neurons were used in the experiments: the U1 group over-expressing U1 snRNA, and the control group. Specifically, two samples of the U1 group, each comprising approximately $1x10^6$ cells, were individually transfected with 50 uL lentivirus solution harboring the U1 snRNA expression plasmids, while the two samples of the control group were transfected with lentivirus solution harboring empty plasmids containing 5-bp polyA sequences. The U1 snRNA expression plasmids were constructed by cloning U1 DNA fragments (comprising the native promoter, coding region and terminator of U1 snRNA) into the pLVX-shRNA1 vectors. The coding region of human U1 snRNA is a 164-bp sequence (RefSeq: NR_004430.2). The lentivirus solution was prepared using the Lenti-X™ HTX Packaging System (Clontech, USA) and the viral titers were adjusted to 10^8 infectious units/ml using Lentivirus Concentration Solution (YEASEN, China). Detailed procedures and methodologies of the transfection experiments have been previously described[15]. Following lentivirus transfection, the four samples were used for library construction and sequencing on an Illumina HiSeq 2000 sequencer. The raw reads of all samples have been deposited in the NCBI SRA database under the project accession number SRP178463. The normalization, DGEA and DSFA of SRP178463 were performed as described in the preceding section. The gene expression matrix and the results of DGEA are available in the NCBI GEO database under the GEO Series accession number GSE124951.

Statements and Declarations

Funding

This work was supported by grants from the National Natural Science Foundation of China (32371153) to Tao Zhang. The funding bodies played no role in the study design, data collection, analysis, interpretation or manuscript writing.

Authors' contributions

Shan Gao conceived the project. Shan Gao and Guangyou Duan supervised this study. Guangyou Duan and Yiyao Zhang performed programming. Jingsong Shi, Shunmei Chen and Chang Liu downloaded, managed and processed the data. Zhi Cheng, Yang Yao, and Dongsheng Wei conducted the experiments. Shan Gao drafted the main manuscript text. Shan Gao and Tao Zhang revised the manuscript.

Acknowledgments

We are grateful for the help from the following faculty members of College of Life Sciences at Nankai University: Wenjun Bu, Dawei Huang, Huaijun Xue and Zhen Ye.

References

1. a, b *Gao S, Ou J, Xiao K. R language and Bioconductor in bioinformatics applications (Chinese Edition). Tianjin: Tianjin Science and Technology Translation Publishing Ltd; 2014.*
2. ^ *Ren Y, et al. Full-length transcriptome sequencing on PacBio platform (in Chinese). Chinese Science Bulletin. 2016; 61(11): 1250-1254.*
3. ^ *Hayer KE, et al. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31(24): 3938-45.*
4. ^ *Li D, et al. Neurodegenerative diseases: a hotbed for splicing defects and the potential therapies. Translational Neurodegeneration. 2021; 10(1): 16.*
5. ^ *Kahles A, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell. 2018; 34(2): 211-224.e6.*
6. a, b, c, d, e *Leung SK, et al. Long-read transcript sequencing identifies differential isoform expression in the entorhinal cortex in a transgenic model of tau pathology. Nature Communications. 2024; 15(1): 6458.*

7. a, b *Berg MG, et al. U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell. 2012; 150(1): 53-64.*
8. ^ *Xu X, et al. Using pan RNA-seq analysis to reveal the ubiquitous existence of 5′ and 3′ end small RNAs. Fro ntiers in Genetics. 2019; 10: 1-11.*
9. ^ *Sarantopoulou D, et al. Comparative evaluation of full-length isoform quantification from RNA-Seq. BMC Bioinformatics. 2021; 22(1): 266.*
10. ^ *Annear DJ, et al. Abundancy of polymorphic CGG repeats in the human genome suggest a broad involvem ent in neurological disease. Scientific Reports. 2021; 11(1): 2515.*
11. ^ *Kaida D, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature. 2010; 468(7324): 664-668.*
12. a, b *Rockenstein EM, et al. Levels and Alternative Splicing of Amyloid β Protein Precursor (APP) Transcripts i n Brains of APP Transgenic Mice and Humans with Alzheimer's Disease (*). Journal of Biological Chemistr y. 1995; 270(47): 28257-28267.*
13. ^ *Ottesen EW, et al. U1 snRNA interactions with deep intronic sequences regulate splicing of multiple exons o f spinal muscular atrophy genes. 2024. 18.*
14. ^ *Zhang M, et al. Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality cont rol. in Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on. 2014. IEEE.*
15. ^ *Cheng Z, et al. Presenilin 1 mutation likely contributes to U1 small nuclear RNA dysregulation and Alzheim er's disease-like symptoms. Neurobiol Aging. 2021; 100: 1-10.*

Declarations

Funding: This work was supported by grants from the National Natural Science Foundation of China (32371153) to Tao Zhang. The funding bodies played no role in the study design, data collection, analysis, interpretation or manuscript writing.

Potential competing interests: No potential competing interests to declare.