PLOS ONE

RESEARCH ARTICLE

Recursive splicing is a rare event in the mouse brain

Sohyun Moon, Ying-Tao Zhao*

Department of Biomedical Sciences, New York Institute of Technology College of Osteopathic Medicine, Old Westbury, New York, United States of America

* yzhao47@nyit.edu

OPEN ACCESS

Citation: Moon S, Zhao Y-T (2022) Recursive splicing is a rare event in the mouse brain. PLoS ONE 17(1): e0263082. https://doi.org/10.1371/journal.pone.0263082

Editor: Alexander F. Palazzo, University of Toronto, CANADA

Received: October 13, 2021

Accepted: January 11, 2022

Published: January 28, 2022

Data Availability Statement: All relevant data can be found in the paper, Supporting Information files, and the following data repositories. The data sets supporting the conclusions of this article are available in the NCBI GEO database with the accession numbers of GSE83474 and GSE90205. The custom code supporting the conclusions of this article is available in the GitHub repository (https://github.com/Jerry-Zhao/RS2020).

Funding: No. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Recursive splicing (RS) is a splicing mechanism to remove long introns from messenger RNA precursors of long genes. Compared to the hundreds of RS events identified in humans and drosophila, only ten RS events have been reported in mice. To further investigate RS in mice, we analyzed RS in the mouse brain, a tissue that is enriched in the expression of long genes. We found that nuclear total RNA sequencing is an efficient approach to investigate RS events. We analyzed 1.15 billion uniquely mapped reads from the nuclear total RNA sequencing data in the mouse cerebral cortex. Unexpectedly, we only identified 20 RS sites, suggesting that RS is a rare event in the mouse brain. We also identified that RS is constitutive between excitatory and inhibitory neurons and between sexes in the mouse cerebral cortex. In addition, we found that the primary sequence context is associated with RS splicing intermediates and distinguishes RS AGGT site from non-RS AGGT sites, indicating the importance of the primary sequence context in RS sites. Moreover, we discovered that cryptic exons may use an RS-like mechanism for splicing. Overall, we provide novel findings about RS in long genes in the mouse brain.

Introduction

Removing introns from the messenger RNA (mRNA) precursors is an essential step of gene expression. This step is affected by the intron lengths. A long intron represents a large RNA molecule to be removed from mRNA precursors. This large molecule poses a challenge for canonical splicing mechanism, which results in a high rate of splicing errors in long introns [1]. Recursive splicing (RS) is a splicing mechanism that removes a long intron into several smaller segments as opposed to in a large single unit [2–12]. RS is untraceable in the mature mRNA, and the direct evidence of RS is the splicing intermediates. However, RS splicing intermediates are unstable, making them difficult to be captured and analyzed. Whole-cell ribosomal RNA-depleted total RNA sequencing (total RNA-seq) and nascent RNA-seq have been used to identify RS splicing intermediates [4, 6, 8–10, 12].

RS has been widely studied in humans. Two earlier studies identified five and nine RS events in humans [4, 6], suggesting that RS might be a rare splicing mechanism for a small group of long introns. Intriguingly, a recent study using nascent RNA-seq identified 342

Competing interests: The authors have declared that no competing interests exist.

candidate RS sites in three human cell lines [10]. Similarly, another study using nascent RNA-seq reported 5,468 RS events in a human cell line [12]. These two studies suggested that RS might be a widely used splicing mechanism in mammals. However, given these extensive studies of RS in humans, only ten RS events have been reported in mice so far. Thus, the extent to which RS is used for intron splicing in mice remains largely unexplored.

Here, we report that nuclear total RNA-seq is enriched for RS splicing intermediates and nascent transcripts, which suggests that it is an efficient approach to investigate RS events. Using nuclear total RNA-seq data generated from the mouse brain, we identified novel RS events, examined the cell-type and sex specificity of RS, and analyzed RS splicing intermediates. We found that RS is a rare process of intron splicing in the mouse brain. We also found that some cryptic exons use an RS-like mechanism for splicing. Together, we provide new findings about RS in long genes in the mouse brain.

Results

Nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates in the mouse brain

Compared to hundreds of RS events identified in humans, only ten RS events have been reported in mice [6]. Because RS tends to occur in long genes, to further investigate RS in mice, we analyzed RS in the mouse brain, a tissue that is enriched in the expression of long genes. To determine whether nascent transcripts and RS splicing intermediates are enriched in nuclear RNA (Fig 1A), we analyzed nuclear ribosomal RNA-depleted total RNA sequencing (total RNA-seq) and whole-cell total RNA-seq data that we recently generated from the mouse

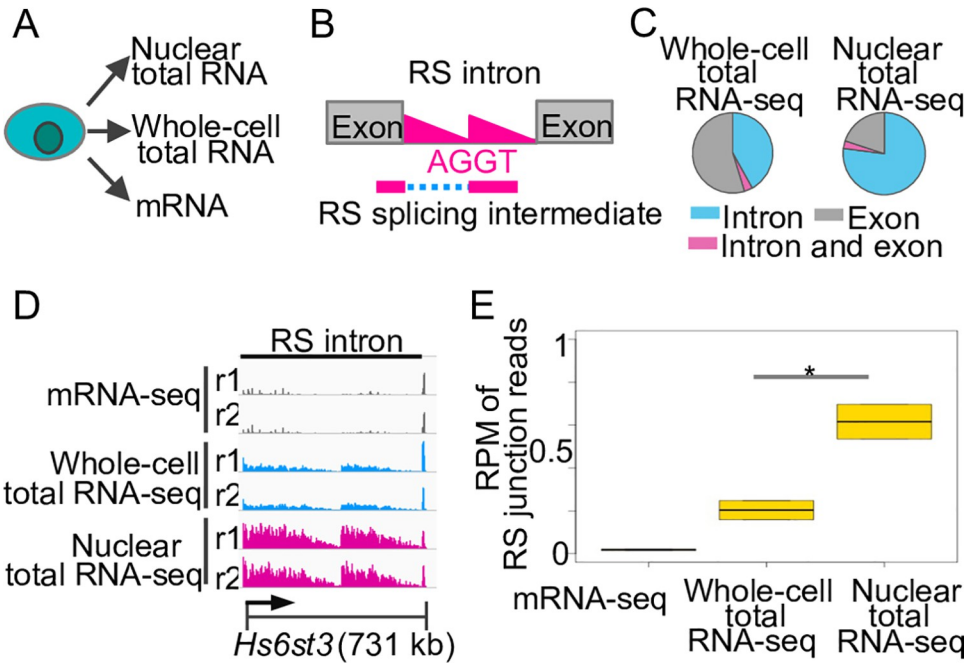

Fig 1. Nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates in mouse brain. (A) Schematic of the isolation of different types of RNA. (B) Schematic of the two features of RS, the saw-tooth pattern (red triangles) and the RS splicing intermediate. (C) Pie charts of loci of uniquely mapped reads in gene regions. (D) Sequencing profile at *Hs6st3*. r1, replicate 1. kb, kilobases. (E) Boxplot of normalized numbers of RS junction reads at *Hs6st3* RS site. RPM, reads per million uniquely mapped reads. *, $P = 0.03$, one-tailed t-test.

https://doi.org/10.1371/journal.pone.0263082.g001

cerebral cortex [13, 14] (S1A Fig). As a control data set, we also analyzed poly(A) enriched messenger RNA-seq (mRNA-seq) data that were generated from the same mouse brain region [15].

Nascent transcripts, which can be indicated by the sequencing reads mapped to introns [16], are important to identify RS events, because RS introns exhibit a saw-tooth pattern of nascent transcript signals from total RNA-seq data [4, 6] (red triangles, Fig 1B). To determine the enrichment of nascent transcripts in nuclear RNA, we calculated the proportions of reads mapped to introns in nuclear total RNA-seq data, whole-cell total RNA-seq data, and mRNA-seq data. We found that 77% of the uniquely mapped reads from the nuclear total RNA-seq data are localized in introns, which is significantly higher than the 41% from the whole-cell total RNA-seq data and the 23% from the mRNA-seq data ($P < 2.2^{-16}$, one-tailed Fisher's Exact Test) (Fig 1C and S1A Fig). In addition, at a known RS intron in *Hs6st3*, we observed a more distinct saw-tooth pattern in the nuclear total RNA-seq data than in the whole-cell total RNA-seq data (Fig 1D). These results suggest that nuclear total RNA is enriched for nascent transcripts.

RS splicing intermediates, which can be indicated by the junction reads spanning the upstream exon and the RS AGGT site (RS junction reads, Fig 1B), is also important to identify RS events. To examine RS splicing intermediates in nuclear RNA, we calculated the numbers of junction reads at a known RS site in *Hs6st3*. We found that the number of RS junction reads at *Hs6st3* is 3-fold higher in nuclear total RNA-seq data than in whole-cell total RNA-seq data (Fig 1E). Together, these results indicate that nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates.

Identify RS sites using nuclear total RNA-seq data

To identify RS sites from nuclear total RNA-seq data, we modified a previously developed pipeline and incorporated additional criteria [6] (Fig 2A). First, we extracted all junction reads spanning the AGGT sites in the mouse genome and then focused on the AGGT sites located in introns that are longer than 1 kb. To exclude unannotated exons from our analyses, we selected AGGT sites that show more than two-fold enrichment of RS junction reads in nuclear total

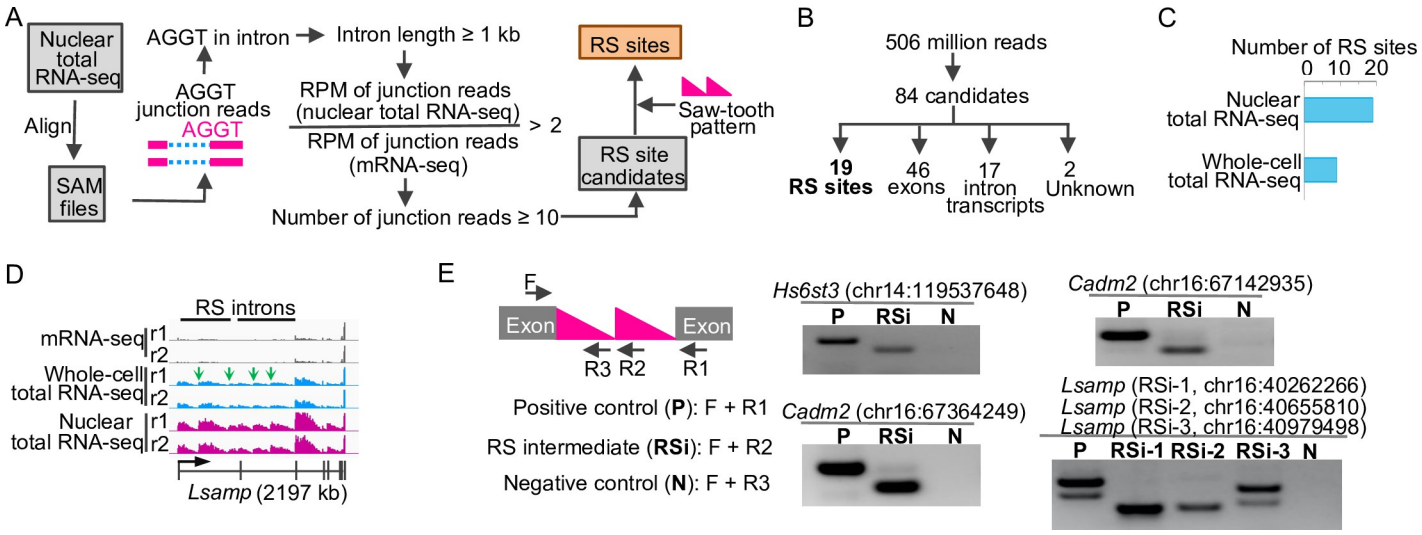

Fig 2. Identify RS sites using nuclear total RNA-seq data. (A) Schematic of the pipeline using nuclear total RNA-seq data to identify RS sites. (B) The 84 RS candidates. (C) Bar plot of RS sites identified using the two sequencing methods in the mouse cerebral cortex. (D) Sequencing profile at *Lsamp* locus. Green arrows indicate the four novel RS sites. (E) Using RT-PCR to detect RS splicing intermediates. Left, schematic of the primer design. Right, RT-PCR and gel results. Note, there are two bands in the positive control and RSi-3 in *Lsamp*, because there is an alternative splicing exon (60 bp) located between RS site 2 and RS site 3 in *Lsamp*.

https://doi.org/10.1371/journal.pone.0263082.g002

RNA-seq than in mRNA-seq. To increase the degree of confidence, we selected sites containing 10 or more RS junction reads as RS site candidates. To enhance the detection of RS junction reads, one of the nuclear total RNA-seq sample was sequenced to extra depth, resulting in 428 million uniquely mapped reads (S1A Fig). Lastly, we visually inspected the candidates to select sites that show saw-tooth patterns in their host introns as RS sites.

RS is a rare event in the mouse brain

By applying the pipeline to the 506 million uniquely mapped reads from our nuclear total RNA-seq data, we identified 84 RS site candidates (S1B Fig). To further refine these candidates, we manually inspected the 84 sites, including the profiles of all sequencing reads and the junction reads at these genomic loci. We found that 19 of the 84 candidates are RS events that show a saw-tooth pattern (Fig 2B, 2C, and S1B Fig). For the other candidates, 46 of them are exons (S1B and S1C Fig) that were not annotated in the Ensembl release 93 database; 17 of them are likely nascent transcripts in introns (S1B and S1C Fig); and 2 of them are unknown sites in the introns of *Etl4* and *Nufip1* (S1B Fig).

The 19 RS sites include all the ten known RS sites in mice [6] (S1D Fig). Nine RS sites (47%) we identified are novel in mice, including the four sites in the introns of *Lsamp* (Fig 2D, green arrows). To validate the identified RS sites, we conducted RT-PCR to detect the RS splicing intermediates for three known RS sites in *Hs6st3* and *Cadm2* and three novel RS sites in *Lsamp* (Fig 2E). Notably, we detected RS splicing intermediates for all the six RS sites (Fig 2E), which independently supported the RS sites that we identified using the nuclear total RNA-seq data.

We next applied the same pipeline to the whole-cell total RNA-seq data and identified nine RS sites (S1D Fig). We found that using nuclear total RNA-seq identified two-fold more RS sites than using whole-cell total RNA-seq (Fig 2C), suggesting that nuclear total RNA-seq is an efficient approach to identify RS events.

To determine whether our criteria were too stringent, we lowered the cutoff of the counts of RS junction reads from 10 to five. We identified 103 additional RS site candidates (S1E Fig). However, none of them exhibited saw-tooth patterns in the host introns, suggesting that lowering the cutoff of the counts of RS junction reads did not identify additional RS sites. Thus, of the 28,382 long introns ($>$ 5 kb) that were actively transcribed in the mouse brain, only less than 0.06% of them show RS. Together, the finding that only 19 RS sites were identified from 506 million uniquely mapped reads indicates that RS is a rare event in the mouse brain.

RS is restricted to the AGGT motif

Although GT is the most highly used motif for 5' splice sites (5'SS), other motifs can also be used for 5'SS. To determine whether RS occurs in AGNN motifs where NN is not GT, we used a pipeline to identify RS sites in AGNN motif. This pipeline is similar to the one in Fig 2A. Briefly, we extracted all junction reads spanning the AGNN sites, focused on the AGNN sites located in introns, selected AGNN sites that show more than two-fold enrichment of RS junction reads in nuclear total RNA-seq than in mRNA-seq, and then selected sites containing 10 or more RS junction reads as RS site candidates. As a result, we identified 405 AGNN RS site candidates (S2 Fig). However, visual inspection revealed that none of them exhibit saw-tooth patterns in the host introns. Thus, these results indicate that RS is restricted to the AGGT motif.

RS is constitutive between cell types and sexes in the mouse cerebral cortex

RS has been shown to be constitutive in Drosophila and cell-type-specific in humans [4, 10]. To determine the cell-type specificity of RS in mouse brain, we focused on two types of

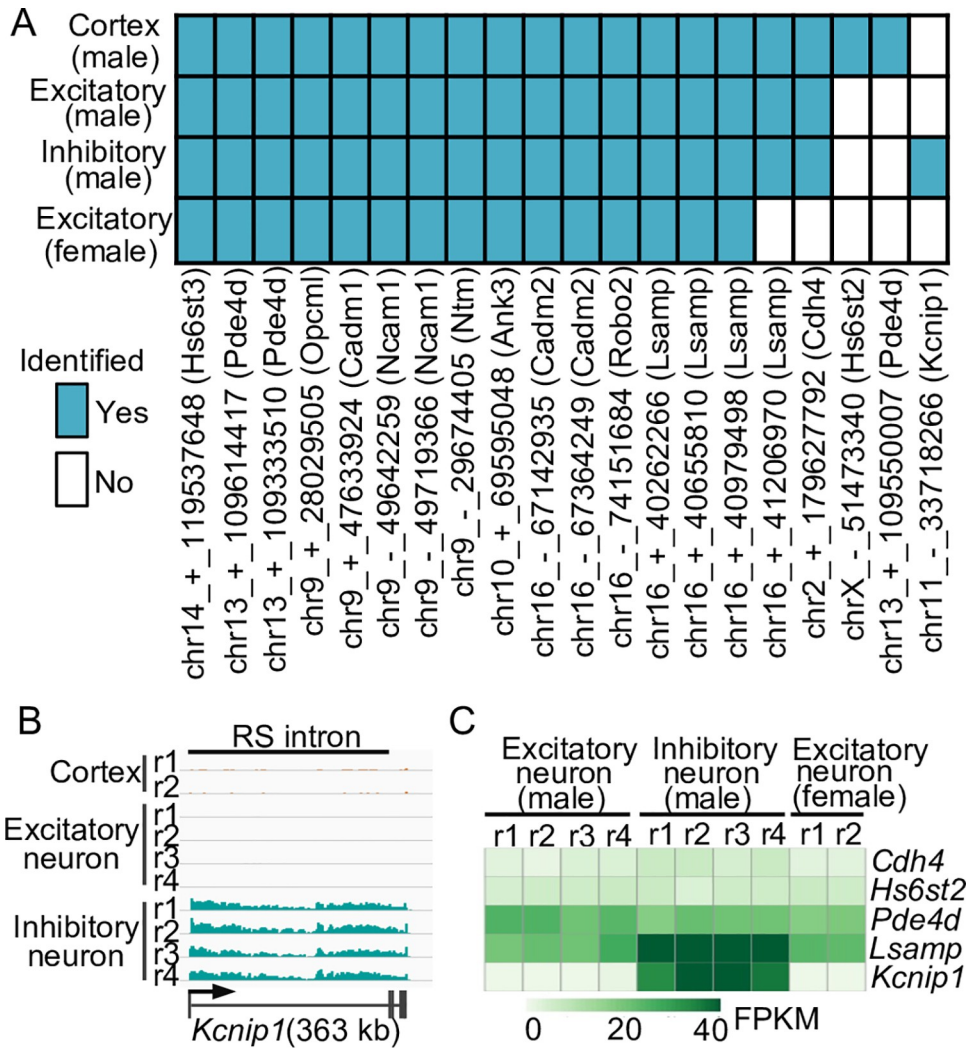

Fig 3. Cell-type and sex specificity of RS in the mouse cortex. (A) Heatmap of RS sites identified in male cortex, male cortical excitatory and inhibitory neurons, and female cortical excitatory neurons. (B) Nuclear total RNA-seq profile (male) at *Kcnip1* locus. (C) Heatmap of expression levels of five genes in three cell types. FPKM, fragment per million uniquely mapped reads per kilobase of exonic region.

https://doi.org/10.1371/journal.pone.0263082.g003

neurons, the excitatory neurons and the inhibitory neurons, which account for 85% and 15% of the neurons in the mouse cerebral cortex [13]. We recently developed a genetic approach to tag and isolate cell-type-specific nuclei from the mouse brain and profiled gene expression in the excitatory and inhibitory neurons using nuclear total RNA-seq [13]. We analyzed the nuclear total RNA-seq data from the two neuronal cell types (S1A Fig). We applied our pipeline to these data and identified 17 RS sites in excitatory neurons and 18 RS sites in inhibitory neurons (Fig 3A and S1D Fig). Notably, all but one of the RS sites are common in both neuronal cell types (Fig 3A). The only RS site unique to inhibitory neurons resides in the *Kcnip1* gene, which is only expressed in the inhibitory neurons (Fig 3B and 3C). Thus, these results indicate that RS is largely constitutive between the two types of neurons in the mouse cerebral cortex.

We recently also profiled gene expression in the excitatory neurons in the cerebral cortex of female mice using nuclear total RNA-seq [13]. To investigate the sex specificity of RS, we

analyzed the nuclear total RNA-seq data of excitatory neurons from the cerebral cortex of female mice [13]. We identified 15 RS sites (S1D Fig), which are all included in the 17 RS sites we identified from male excitatory neurons (Fig 3A). The remaining two RS sites are unlikely male specific, because we also identified seven and five junction reads for them in the female data (S1D Fig), although they failed to pass our criteria of 10 junction reads. Together, these results indicate that RS is constitutive between male and female excitatory neurons in the mouse cerebral cortex.

Characteristics of RS sites

In consistent with previous reports, RS sites are highly conserved (S3A Fig), are located in long introns (> 51 kb, S3B Fig), are specific to long genes (> 294 kb), prefer the first two introns (S3C Fig, 75% in first intron and 25% in second intron), and are located in genes that are specifically expressed in the brain (S3D Fig). Furthermore, compared to non-RS AGGT sites, RS AGGT sites are enriched for AGGTAAGT motif that complements with the 5' conserved sequence of U1 snRNA (S3E–S3G Fig), are enriched for thymine and cytosine in the 20 nt regions upstream of RS AGGT sites (polypyrimidine tract, S3H Fig), and show higher 3' splice site scores when examined using MaxEntScan [17] (S3I Fig).

Splicing intermediates linking RS exons to the downstream annotated exons

RS uses exon definition mechanism [6, 8], thus suggesting a possibility of splicing intermediates linking RS exons to the downstream annotated exons (downstream splicing intermediates, Fig 4A). Although Sibley et al. reported that RS exons were not detectable in mRNA transcripts and were included only after blocking the 5'SS using antisense oligonucleotide [6], the extent to which the downstream splicing intermediates exist in physiological conditions remains largely unexplored.

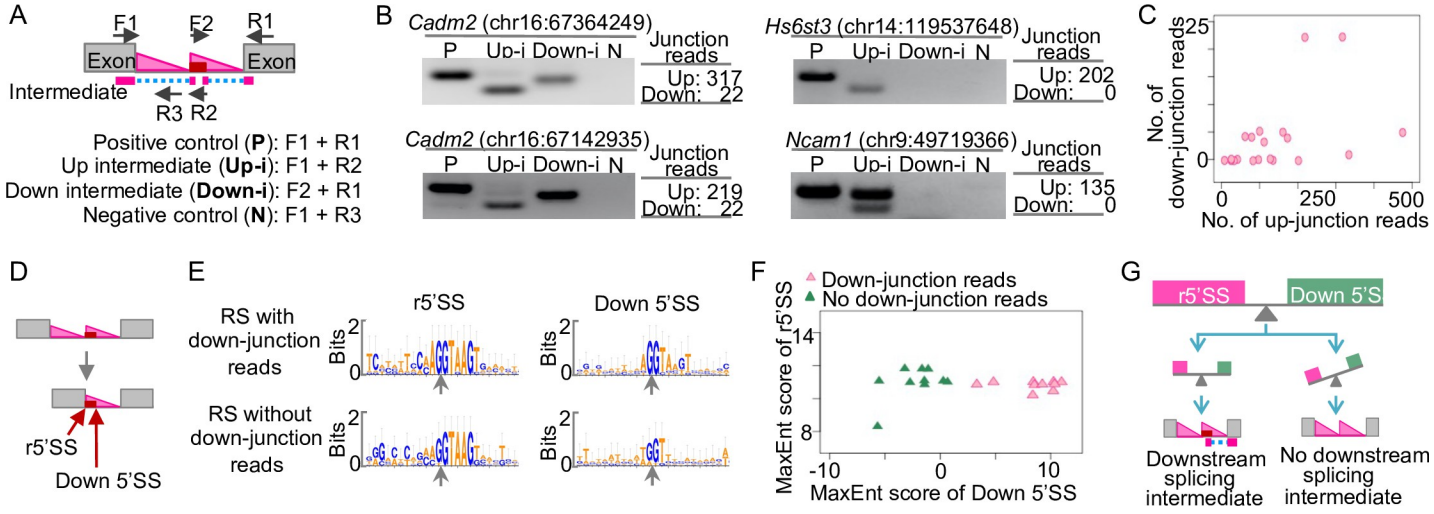

Fig 4. Downstream RS splicing intermediates. (A) Schematic of the up- and down-junction reads that represent the upstream and downstream RS splicing intermediates and the primer design. (B) Validation of the upstream (Up-i) and downstream (Down-i) RS splicing intermediates using RT-PCR. P, positive control. N, negative control. Left, RT-PCR and gel results. Right, numbers of junction reads from the nuclear total RNA-seq data. (C) Scatter plot of the numbers of up- and down-junction reads of the 20 RS sites. (D) The reconstituted 5'SS (r5'SS) and the downstream 5'SS (Down 5'SS). (E) Sequence logos of the r5'SS and the Down 5'SS of RS sites with or without down-junction reads. (F) Scatter plot of the MaxEnt scores of r5'SS and Down 5'SS of RS sites. (G) Model of the 5'SS strengths and the downstream splicing intermediates.

https://doi.org/10.1371/journal.pone.0263082.g004

To identify downstream splicing intermediates of RS sites, we developed a computational pipeline to identify junction reads that link RS exons to downstream annotated exons (down-junction reads, Fig 4A). Briefly, we obtained all the junction reads from our nuclear total RNA-seq data and then scanned the 350 nt region downstream of the RS AGGT site for junction reads that link to the downstream exons. We found down-junction reads for 10 RS sites, but not for the other 10 sites (S4A Fig). To validate the findings of down-junction reads, we conducted RT-PCR to detect the downstream splicing intermediates for four RS sites (Fig 4B). Notably, for the two RS sites that show both up- and down-junction reads, we detected both upstream and downstream splicing intermediates (Fig 4B, left panel). However, for the two RS sites that only show up-junction reads, we only detected the upstream splicing intermediates, not the downstream splicing intermediates (Fig 4B, right panel). Thus, the RT-PCR results are consistent with the results from the nuclear total RNA-seq. Together, these results indicate that ten RS sites have downstream splicing intermediates, while the other ten sites do not have downstream splicing intermediates.

Downstream splicing intermediates are associated with high strengths of the 5' splice sites at the 3' end of the RS exons

To understand why some RS sites do not show downstream splicing intermediates, we first asked whether these RS sites also show low abundance of upstream splicing intermediates. We analyzed the numbers of up- and down-junction reads for the 20 RS sites. However, we found no positive correlation between the two (Fig 4C), suggesting that the absence of downstream splicing intermediates for the 10 RS sites is not associated with the low abundance of upstream splicing intermediates.

Previous studies have showed that the competition between the two 5'SS was associated with the inclusion of the RS exons in mature transcripts [6, 8, 10]. Thus, another hypothesis is that the 5'SS at the two ends of the RS exons, the reconstituted 5'SS (r5'SS) after the first step of splicing and the downstream 5'SS (Down 5'SS) (Fig 4D), affect the existence of the downstream splicing intermediates. To test this hypothesis, we quantified the strengths of r5'SS and Down 5'SS using MaxEntScan [17]. We found that RS sites with down-junction reads show higher MaxEnt scores of Down 5'SS compared to RS sites without down-junction reads ($P = 4.28^{-9}$, one-tailed t-test) (Fig 4E, 4F, and S4A Fig). By contrast, the scores of r5'SS are comparable between the two groups (Fig 4E, 4F, and S4A Fig). Together, these results support a model that the downstream splicing intermediates of RS sites are likely associated with the high strengths of 5'SS at the 3' end of the RS exons (Fig 4G).

Cryptic exons use a RS-like splicing mechanism

During our analysis of RS site candidates, we found genomic loci that show exon-like splicing patterns (Fig 2B). Some of these loci were later annotated as exons in Ensembl database, while the others are not. We termed these unannotated genomic loci as RS-like cryptic exons (Fig 5A). To systematically identify RS-like cryptic exons, we developed a computational pipeline (S4B Fig). To exclude unannotated exons from our downstream analysis, we used a criterion that the normalized count of up-junction reads in nuclear total RNA-seq should be higher than that in mRNA-seq. By applying this pipeline to our nuclear total RNA-seq data, we identified 22 RS-like cryptic exons in the introns of 21 long genes (S4C Fig). To validate the RS-like cryptic exons, we conducted RT-PCR to detect splicing intermediates for three RS-like cryptic exons. We confirmed the upstream and downstream splicing intermediates for all the three RS-like cryptic exons (Fig 5B). Notably, we found that although the cryptic exons in *Lrrc4c* and *Magi1* were not included in the mature mRNA, the cryptic exon in *Nova1* was included in a

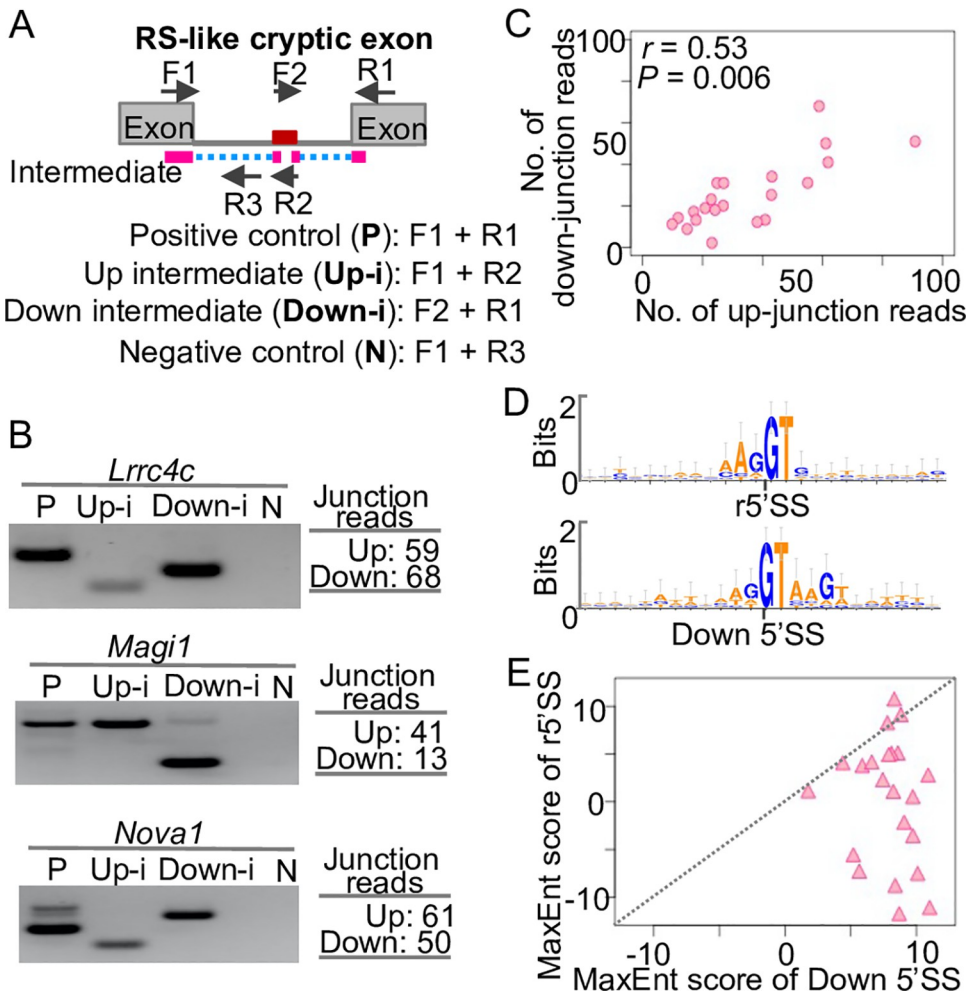

Fig 5. Cryptic exons use a RS-like splicing mechanism. (A) Schematic of RS-like cryptic exons and the primer design. (B) Validation of the upstream (Up-i) and downstream (Down-i) splicing intermediates using RT-PCR. P, positive control. N, negative control. Left, RT-PCR and gel results. Right, numbers of junction reads from the nuclear total RNA-seq data. (C) Scatter plot of the numbers of up- and down-junction reads of the RS-like cryptic exons. (D) Sequence logos of the r5'SS and Down 5'SS of the RS-like cryptic exons. (E) Scatter plot of the MaxEnt scores of the r5'SS and Down 5'SS of the RS-like cryptic exons.

https://doi.org/10.1371/journal.pone.0263082.g005

part of the mature mRNA (Fig 5B), suggesting that RS-like cryptic exons could be included in the mature mRNA.

To identify features of RS-like cryptic exons, we analyzed the correlation between the up- and down-junction reads for them. We found that RS-like cryptic exons exhibit a positive correlation between the numbers of up- and down-junction reads (Pearson correlation coefficient $r = 0.53$, *P-value* = 0.01, Fig 5C), which contrasts with the results of RS sites (Fig 4C). To investigate the strengths of the 5'SS at the two ends of RS-like cryptic exons, we examined the sequence compositions of the r5'SS and Down 5'SS using WebLogo [18]. We found that the RS-like cryptic exons show an enrichment of 5'SS motif (AGGTAAGT) at the Down 5'SS but not at the r5'SS (Fig 5D). Furthermore, we quantified the strengths of r5'SS and Down 5'SS using MaxEntScan [17] and found that the MaxEnt scores of r5'SS are significantly lower than that of Down 5'SS in RS-like cryptic exons (*P-value* < 0.0001, one-tailed t-test, Fig 5E). One exception is the RS-like cryptic exon in *Magi1*, which shows a much higher MaxEnt score at

the r5'SS than the Down 5'SS (Fig 5E and S4C Fig). Notably, the *Magi1* intron exhibits a weak saw-tooth pattern (S4D Fig), which failed to pass the stringent criteria in our identification of RS sites. Taken together, these results demonstrate that some cryptic exons may use an RS-like mechanism for splicing.

Discussion

In this study, we demonstrated that RS is a rare event in the mouse brain and that nuclear total RNA-seq is an efficient approach to investigate RS events. We developed a novel pipeline to identify RS sites, determined the cell-type and sex specificity of RS, and characterized the genomic features of RS sites. Through analysis of the primary sequences, we found that RS splicing intermediates are associated with the primary sequence context. We also discovered that some cryptic exons use an RS-like mechanism for splicing. The pipeline we developed can be applied to identify RS using other types of RNA-seq data sets from different tissues and species, which will be a useful tool for the RNA community.

The extent to which RS is used for intron splicing in mammals remains inconclusive. Here we analyzed high-depth nuclear total RNA-seq data and identified 20 RS sites in mice, which is consistent with previous studies [4, 6] that RS is a rare splicing process for a small group of long introns in the mammalian genome. We found that whether or not including the saw-tooth patterns to define RS sites might be the major reason for the differences of the reported numbers of RS events [4, 6, 10, 12]. For example, we identified 65 AGGT sites and 405 AGNN sites that have 10 or more junction reads, but they did not show saw-tooth patterns in the host introns. Therefore, we did not consider them as RS sites. Notably, the relationship between saw-tooth patterns and RS events remains unclear. Further studies are needed to investigate whether the saw-tooth pattern should be a required criterion to define an RS event.

Long introns exhibit a high rate of creating new exons during evolution [19], but the underlying mechanisms are not fully understood. Our discovery of RS-like cryptic exons indicates that long introns may acquire novel exons via the RS-like cryptic exons. This is supported by the findings that more than 6000 human annotated exons are RS-like exons [20]. In addition, the numbers of RS sites in *Drosophila melanogaster* are about 15 times more than that in humans [4, 6, 8–10], but the numbers of RS-like annotated exons in *Drosophila* are 2~100 times less than that in humans [6, 8, 20]. Thus, future studies are needed to investigate RS, RS-like cryptic exons, and RS-like annotated exons in evolutionarily distinct species to determine their associations during evolution.

RS genes in mice are long genes ($>$ 294 kb). Long genes have been shown to exhibit a series of unique genomic and epigenomic features. For instance, we have found that actively transcribed long genes in the mouse brain harbor broad enhancer-like chromatin domains and show high levels of transcription initiation frequency and chromatin-chromatin interactions [14]. Thus, RS could be another unique mechanism that is utilized by long genes. Further studies are needed to illustrate the mechanisms by which only a subset of long genes uses the RS mechanism for splicing.

RS genes are genetically linked to various human brain disorders. For example, *ANK3* encodes ankyrin-G and is linked to autism spectrum disorders, attention deficit hyperactivity disorder, intellectual disability, and bipolar disorder [21–24]. Also, *NTM* encodes neurotrimin and is linked to autism spectrum disorders and attention deficit hyperactivity disorders [25, 26]. The *PDE4D* encodes phosphodiesterase 4D and is linked to schizophrenia, psychosis, acrodysostosis, and neuroticism [27–29]. Notably, PDE4D Inhibitors are in clinical trials for the treatment of Alzheimer's disease and Fragile X syndrome [30, 31]. Given that the disruption of the RS process interfered with the RS gene function and caused abnormality in the

central nervous system [8], further investigation will be necessary to illuminate whether RS mechanism contributes to the pathophysiology of these human brain disorders.

Materials and methods

Animals

The C57BL/6 mice were obtained from the Jackson Laboratory. Mice were maintained in the same genetic background and were kept on a regular 12-hour light/12-hour dark cycle. Animal studies were carried out in accordance with the National Institutes of Health's Guide for the Care and Use of Laboratory Animals recommendations. The study protocol was approved by the Institutional Animal Care and Use Committee (IACUC) of the New York Institute of Technology. The study is reported in accordance with the ARRIVE guidelines (https://arriveguidelines.org).

RNA isolation and RT-PCR

The brain was dissected from 8-week-old mice. Total RNA was extracted from the brain tissues using TRIzol reagent (Invitrogen, #15596026) according to the manufacturer's instructions. The integrity and quality of the isolated RNA were determined by the Nanodrop and the Bioanalyzer. The Prime script reverse transcript kit (Takara, #RR037A) was used to synthesize the cDNA. The RT-PCR experiments were conducted using the GoTaq G2 Hot Start Master mix (Promega #M7422). The PCR primers were listed in the S5 Fig. The RT-PCR were conducted using the following setting: 94˚C for 3 minutes, followed by 32–40 cycles of thermocycling (94˚C for 30 seconds, 52˚C-56˚C for 30 seconds, 72˚C for 30 seconds), and 72˚C for 3 minutes.

Statistical analysis

All statistical analyses were performed in the R software version 3.6.1 (https://www.r-project.org/).

Nuclear total RNA-seq data analysis

The analysis pipeline was described previously [13, 14, 32, 33]. In brief, raw data in sra files were downloaded from the EBI European Nucleotide Archive database [34] using the accession numbers listed in S1 Fig. The fastq-dump.2.9.6 of NCBI SRA ToolKit was used to extract the FASTQ files using the parameter of "—split-3". STAR [35] was used to map the FASTQ raw reads into mouse mm10 genome using the parameters of "—runThreadN 40—outFilterMultimapNmax 1—outFilterMismatchNmax 3". The samtools view [36] was used to convert sam files into bam files. The samtools sort was used to sort the bam files. The samtools index was used to index the sorted bam files. The bamCoverage [37] was used to convert the sorted bam files into strand-specific bigwig files. The bamCoverage parameters that were used included "—filterRNAstrand forward—binSize 1 -p 40 -o" for plus strand and "—filterRNAstrand reverse—binSize 1 -p 40 -o" for minus strand.

Sequencing data visualization

All sequencing data, including the bigwig files and bam files, were visualized in the IGV_2.8.2 genome browser [38].

Genome annotation

The gtf file of mouse genome annotation was downloaded from the Ensembl release 93 [39].

Junction reads

A read pair is considered as a junction read if its CIGAR in sam files contains "N". Junction reads were extracted from sam files and were saved into a junction-read-specific sam files. These sam files were further converted into bam files using samtools. The junction-read-specific bam files were loaded into IGV for visualization.

Junction reads spanning AGGT sites

All AGGT sites (20,403,114) in the mouse mm10 genome were identified, and only AGGT sites (4,767,575) located in the gene sense regions were kept for further analysis. The AGGT sites that were kept were used to screen the junction-read-specific sam files. The numbers of junction reads spanning each AGGT site (joining the upstream exon and sequences following GT) were counted. The counts of junction reads were further normalized to the sequencing depth to obtain the RPM values.

Pipeline to identify RS sites

The schematic of this pipeline is shown in Fig 2A. Briefly, AGGT sites located in introns longer than or equal to 1 kb were extracted. Sites that showed a larger than two-fold RPM value of junction reads in total RNA-seq data than in mRNA-seq data were kept for downstream analyses. The counts of junction reads of biological replicates were merged, and AGGT sites that contained 10 or more junction reads were identified as RS site candidates. The RS site candidates were further refined as RS sites if the host intron showed a clear saw-tooth pattern by visual inspection.

FPKM of nuclear total RNA-seq data

The number of reads mapped to the exonic regions of each gene were calculated to get the raw counts. The raw counts were then normalized to the exon lengths of that gene and to the sequencing depth of that data set to get the FPKM values.

Phylogenetic p-value (phyloP) scores

The phyloP scores, which were calculated by the PHAST package for multiple alignments of 59 vertebrate genomes to the mouse genome, were obtained from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phyloP60way/).

Gene expression profiles in 22 mouse tissues

The expression profiles (FPKM values) of RS genes in 22 mouse tissues were obtained from the LongGeneDB database (https://longgenedb.com).

WebLogo analysis

WebLogo 3 [18] (http://weblogo.threeplusone.com) was used to perform the sequence logo analysis. The Output Format was chosen as "PNG (high res.)", and the Stacks per Line was set to "80". The default values were used for other parameters.

MaxEntScan 3'SS analysis

The 3'SS scores were calculated by MaxEntScan::score3ss (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html) [17]. The input sequences were composed of the 18 nt region upstream of the AGGT, the AGGT motif, and the one nucleotide following AGGT (18nt + AGGT + 1nt). The three models—Maximum Entropy Model, First-order Markov Model, and Weight Matrix Model—were selected. The MaxEnt scores were used as the 3'SS scores.

Reconstituted 5' splice sites (r5'SS)

The r5'SS sequences were composed of the last 30 nucleotides of the upstream exon, the GT motif, and the 20 nucleotides following AGGT (30nt + GT + 20nt).

MaxEntScan 5'SS analysis

The 5'SS scores were calculated by MaxEntScan::score5ss (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html) [17]. The input sequences for 5'SS were composed of three nucleotides before the GT, the GT motif, and the four nucleotides following AGGT (3nt + GT + 4nt). The input sequences of r5'SS and Down 5'SS were listed in S4A and S4C Fig.

Pipeline to identify RS-like cryptic exons

The schematic of this pipeline is shown in S4B Fig. Briefly, AGGT sites located in introns longer than or equal to 50 kb were extracted. The AGGT sites that showed a larger RPM value of junction reads in total RNA-seq data than in mRNA-seq data were kept for downstream analyses. The counts of the junction reads of the biological replicates were merged. AGGT sites contained 10 or more up-junction reads and two or more down-junction reads were identified as candidates of RS-like cryptic exons. The RS-like cryptic exon candidates were further refined as RS-like cryptic exons if the host intron showed an exon-like but not saw-tooth like pattern.

Supporting information

S1 Fig. Novel RS sites. (A) The mapping statistics and access numbers of RNA-seq data utilized in this study. (B) The loci and numbers of junction reads (#) of the 84 RS site candidates. (C) Snapshots of the junction reads at the *Huwe1* and *Lmbrd1* RS candidate sites. (D) Heatmap of numbers of RS junction reads at each RS site in different total RNA-seq data sets. A green box indicates that the RS site was identified in that data set. Note, the enrichment of junction reads in whole cell data compared to that in mRNA-seq data is less than 2-fold for *Ank3*. (E) The loci and read numbers (#) of the top 50 additional RS site candidates when lowering the cutoff of junction read count from 10 to 5.
(PDF)

S2 Fig. The 405 AGNN RS site candidates. #, number of junction reads.
(PDF)

S3 Fig. Characteristics of RS sites. (A) Heatmap of phyloP score of RS sites and the flanking regions. (B) Boxplot of lengths of RS introns and introns transcribed in the mouse cortex. ***, $P < 0.0001$, one-tailed t-test. (C) Pie chart of locations of RS introns in host genes. (D) Heatmap of expression levels of RS genes in 22 mouse tissues. (E) The sequence motifs, nucleotide percentages, and 3'SS MaxEnt scores of the 20 RS sites. (F) Sequence logos of the 64 nt regions surrounding the 2640 non-RS AGGT sites and the 20 RS AGGT sites. (G) Schematic of the sequence base pairing between the AGGTAAGT motif and U1 snRNA. (H) Boxplots of the

percentages of nucleotides in the 20 nt region upstream of the 2640 non-RS AGGT sites and the 20 RS AGGT sites. (I) Boxplot of MaxEnt 3' splice site (3'SS) scores of the 20 RS AGGT sites and the 2640 non-RS AGGT sites. ***, $P < 0.0001$, one-tailed t-test.
(PDF)

S4 Fig. RS exons and RS-like cryptic exons. (A) The genomic loci, numbers of junction reads, 5'SS sequences, and 5'SS MaxEnt scores of RS exons. (B) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS-like cryptic exons. (C) The genomic loci, numbers of junction reads, 5'SS sequences, and 5'SS MaxEnt scores of RS-like cryptic exons. (D) Sequencing profile at *Magi1* locus. Green arrows indicate the putative RS AGGT loci.
(PDF)

S5 Fig. RT-PCR primer sequences for the RS sites (top table) and RS-like cryptic exons (bottom table). *Cadm2* RS1: chr16:67364249. *Cadm2* RS2: chr16:67142935. *Lsamp* RS1: chr16:40262266. *Lsamp* RS2: chr16:40655810. *Lsamp* RS3: chr16:40979498.
(PDF)

S6 Fig. The full-length gels for plots in Fig 2.
(PDF)

S7 Fig. The full-length gels for plots in Fig 4.
(PDF)

S8 Fig. The full-length gels for plots in Fig 5.
(PDF)

Acknowledgments

We thank members of the Zhao Laboratory for helpful discussions and comments on the manuscript. We thank the Center for Biomedical Innovation at the New York Institute of Technology College of Osteopathic Medicine for support.

Author Contributions

Conceptualization: Ying-Tao Zhao.

Data curation: Sohyun Moon, Ying-Tao Zhao.

Formal analysis: Sohyun Moon, Ying-Tao Zhao.

Funding acquisition: Ying-Tao Zhao.

Investigation: Sohyun Moon, Ying-Tao Zhao.

Methodology: Ying-Tao Zhao.

Project administration: Ying-Tao Zhao.

Resources: Sohyun Moon, Ying-Tao Zhao.

Software: Ying-Tao Zhao.

Supervision: Ying-Tao Zhao.

Validation: Ying-Tao Zhao.

Visualization: Sohyun Moon, Ying-Tao Zhao.

Writing – original draft: Ying-Tao Zhao.

Writing – review & editing: Sohyun Moon, Ying-Tao Zhao.

References

1. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 2010; 6(12):e1001236. https://doi.org/10.1371/journal.pgen.1001236 PMID: 21151575
2. Hatton AR, Subramaniam V, Lopez AJ. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell. 1998; 2(6):787–96. https://doi.org/10.1016/s1097-2765(00)80293-2 PMID: 9885566
3. Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics. 2005; 170(2):661–74. https://doi.org/10.1534/genetics.104.039701 PMID: 15802507
4. Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, et al. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature. 2015; 521(7552):376–9. https://doi.org/10.1038/nature14475 PMID: 25970244
5. Kelly S, Georgomanolis T, Zirkel A, Diermeier S, O'Reilly D, Murphy S, et al. Splicing of many human genes involves sites embedded within introns. Nucleic Acids Res. 2015; 43(9):4721–32. https://doi.org/10.1093/nar/gkv386 PMID: 25897131
6. Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, et al. Recursive splicing in long vertebrate genes. Nature. 2015; 521(7552):371–5. https://doi.org/10.1038/nature14466 PMID: 25970246
7. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018; 9(1):619. https://doi.org/10.1038/s41467-018-02866-0 PMID: 29434199
8. Joseph B, Kondo S, Lai EC. Short cryptic exons mediate recursive splicing in Drosophila. Nat Struct Mol Biol. 2018; 25(5):365–71. https://doi.org/10.1038/s41594-018-0052-6 PMID: 29632374
9. Pai AA, Paggi JM, Yan P, Adelman K, Burge CB. Numerous recursive sites contribute to accuracy of splicing in long introns in flies. PLoS Genet. 2018; 14(8):e1007588. https://doi.org/10.1371/journal.pgen.1007588 PMID: 30148878
10. Zhang XO, Fu Y, Mou H, Xue W, Weng Z. The temporal landscape of recursive splicing during Pol II transcription elongation in human cells. PLoS Genet. 2018; 14(8):e1007579. https://doi.org/10.1371/journal.pgen.1007579 PMID: 30148885
11. Joseph B, Lai EC. The exon junction complex and intron removal prevent re-splicing of mRNA. PLoS Genet. 2021; 17(5):e1009563. https://doi.org/10.1371/journal.pgen.1009563 PMID: 34033644
12. Wan Y, Anastasakis DG, Rodriguez J, Palangat M, Gudla P, Zaki G, et al. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. Cell. 2021; 184(11):2878–95 e20. https://doi.org/10.1016/j.cell.2021.04.012 PMID: 33979654
13. Johnson BS, Zhao YT, Fasolino M, Lamonica JM, Kim YJ, Georgakilas G, et al. Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome. Nat Med. 2017; 23(10):1203–14. https://doi.org/10.1038/nm.4406 PMID: 28920956
14. Zhao YT, Kwon DY, Johnson BS, Fasolino M, Lamonica JM, Kim YJ, et al. Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains. Genome Res. 2018; 28(7):933–42. https://doi.org/10.1101/gr.233775.117 PMID: 29848492
15. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015; 6:5903. https://doi.org/10.1038/ncomms6903 PMID: 25582907
16. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol. 2011; 18(12):1435–40. https://doi.org/10.1038/nsmb.2143 PMID: 22056773
17. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004; 11(2–3):377–94. https://doi.org/10.1089/1066527041410418 PMID: 15285897
18. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14(6):1188–90. https://doi.org/10.1101/gr.849004 PMID: 15173120
19. Roy M, Kim N, Xing Y, Lee C. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA. 2008; 14(11):2261–73. https://doi.org/10.1261/rna.1024908 PMID: 18796579
20. Blazquez L, Emmett W, Faraway R, Pineda JMB, Bajew S, Gohr A, et al. Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. Mol Cell. 2018; 72(3):496–509 e9. https://doi.org/10.1016/j.molcel.2018.09.033 PMID: 30388411

21. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 2008; 40(9):1056–8. https://doi.org/10.1038/ng.209 PMID: 18711365

22. Schulze TG, Detera-Wadleigh SD, Akula N, Gupta A, Kassem L, Steele J, et al. Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. Mol Psychiatry. 2009; 14(5):487–91. https://doi.org/10.1038/mp.2008.134 PMID: 19088739

23. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012; 485 (7397):237–41. https://doi.org/10.1038/nature10945 PMID: 22495306

24. Iqbal Z, Vandeweyer G, van der Voet M, Waryah AM, Zahoor MY, Besseling JA, et al. Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. Hum Mol Genet. 2013; 22(10):1960–70. https://doi.org/10.1093/hmg/ddt043 PMID: 23390136

25. Maruani A, Huguet G, Beggiato A, ElMaleh M, Toro R, Leblond CS, et al. 11q24.2–25 micro-rearrangements in autism spectrum disorders: Relation to brain structures. Am J Med Genet A. 2015; 167A (12):3019–30. https://doi.org/10.1002/ajmg.a.37345 PMID: 26334118

26. Brevik EJ, van Donkelaar MM, Weber H, Sanchez-Mora C, Jacob C, Rivero O, et al. Genome-wide analyses of aggressiveness in attention-deficit hyperactivity disorder. Am J Med Genet B Neuropsychiatr Genet. 2016; 171(5):733–47. https://doi.org/10.1002/ajmg.b.32434 PMID: 27021288

27. Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, et al. A whole genome association study of neuroticism using DNA pooling. Mol Psychiatry. 2008; 13(3):302–12. https://doi.org/10.1038/sj.mp.4002048 PMID: 17667963

28. Lee H, Graham JM Jr., Rimoin DL, Lachman RS, Krejci P, Tompson SW, et al. Exome sequencing identifies PDE4D mutations in acrodysostosis. Am J Hum Genet. 2012; 90(4):746–51. https://doi.org/10.1016/j.ajhg.2012.03.004 PMID: 22464252

29. Sinha V, Ukkola-Vuoti L, Ortega-Alonso A, Torniainen-Holm M, Therman S, Tuulio-Henriksson A, et al. Variants in regulatory elements of PDE4D associate with major mental illness in the Finnish population. Mol Psychiatry. 2019. https://doi.org/10.1038/s41380-019-0429-x PMID: 31138891

30. Gurney ME, Nugent RA, Mo X, Sindac JA, Hagen TJ, Fox D, 3rd, et al. Design and Synthesis of Selective Phosphodiesterase 4D (PDE4D) Allosteric Inhibitors for the Treatment of Fragile X Syndrome and Other Brain Disorders. J Med Chem. 2019; 62(10):4884–901. https://doi.org/10.1021/acs.jmedchem.9b00193 PMID: 31013090

31. Pan T, Xie S, Zhou Y, Hu J, Luo H, Li X, et al. Dual functional cholinesterase and PDE4D inhibitors for the treatment of Alzheimer's disease: Design, synthesis and evaluation of tacrine-pyrazolo[3,4-b]pyridine hybrids. Bioorg Med Chem Lett. 2019; 29(16):2150–2. https://doi.org/10.1016/j.bmcl.2019.06.056 PMID: 31281020

32. Zhao YT, Fasolino M, Zhou Z. Locus- and cell type-specific epigenetic switching during cellular differentiation in mammals. Front Biol (Beijing). 2016; 11(4):311–22. https://doi.org/10.1007/s11515-016-1411-5 PMID: 28261266

33. Moon S, Zhao YT. Spatial, temporal, and cell-type-specific expression profiles of genes encoding heparan sulfate biosynthesis enzymes and proteoglycan core proteins. Glycobiology. 2021.

34. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, et al. The European Nucleotide Archive in 2019. Nucleic Acids Res. 2020; 48(D1):D70–D6. https://doi.org/10.1093/nar/gkz1063 PMID: 31722421

35. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015; 51:11 4 1–4 9. https://doi.org/10.1002/0471250953.bi1114s51 PMID: 26334920

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

37. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016; 44(W1):W160–5. https://doi.org/10.1093/nar/gkw257 PMID: 27079975

38. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29(1):24–6. https://doi.org/10.1038/nbt.1754 PMID: 21221095

39. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020; 48(D1):D682–D8. https://doi.org/10.1093/nar/gkz966 PMID: 31691826