**Primary sequence context defines recursive splicing in long genes**

Jacob Vazquez[1], Ying-Tao Zhao[1*]

[1]Department of Biomedical Sciences, New York Institute of Technology College of Osteopathic Medicine, Old Westbury, NY, 11568, USA.

Lead Contact: Ying-Tao Zhao

*Correspondence: yzhao47@nyit.edu

## Summary

Recursive splicing (RS) is a mechanism to excise long introns from messenger RNA precursors. We focused on nuclear RNA, which is enriched for RS splicing intermediates and nascent transcripts, to investigate RS in the mouse brain. We identified new RS sites and discovered that RS is constitutive between excitatory and inhibitory neurons and between sexes in the mouse cerebral cortex. Moreover, we found that the primary sequence context, including the U1 snRNA binding site, the polypyrimidine tract, and a strong 3' splice site, determines RS sites in long introns. We also uncovered a new type of exon-like RS events termed exonicRS that may represent an intermediate stage of RS sites evolving into annotated exons. Furthermore, the strengths of 5' splice sites are able to classify RS events into RS sites and exonicRS. Overall, these findings provide mechanistic insights into the splicing and evolution of long genes.

## Keywords

Recursive splicing, Long intron, Long gene, RS site, RS exon, Nuclear total RNA-seq

## Introduction

Splicing is a biological process in which introns are excised from messenger RNA (mRNA) precursors. This process is impacted by the length of an intron because a longer intron requires a longer time to be transcribed and results in a larger RNA molecule to be removed than a short intron does. The human genome contains 4023 introns that are

each longer than 50 kilobases (kb). These introns reside in 2193 long genes that encodes proteins that are essential for many biological processes. Notably, long introns are associated with high rates of splicing errors(Pickrell et al., 2010), and splicing errors are implicated in various human diseases(Faustino and Cooper, 2003; Scotti and Swanson, 2016). Therefore, it is crucial to maintain the splicing fidelity of long introns. To facilitate the splicing of long introns in long genes, there are specific mechanisms that can be used, such as recursive splicing (RS).

The RS is a splicing mechanism that is specific to long introns(Burnette et al., 2005; Cook-Andersen and Wilkinson, 2015; Duff et al., 2015; Hatton et al., 1998; Hayashi et al., 2018; Joseph et al., 2018; Kelly et al., 2015; Pai et al., 2018; Sibley et al., 2015; Zhang et al., 2018). This mechanism excises a long intron into several smaller segments as opposed to in a large single unit. Unlike canonical splicing, RS is untraceable in mature mRNAs. Thus, the direct evidence of RS is the splicing intermediates. However, RS splicing intermediates are unstable, making them difficult to be captured and analyzed. These technical challenges have significantly hindered research on RS. Since the first report of RS in 1998, only three RS-containing genes (RS genes) were identified in 17 years(Burnette et al., 2005; Hatton et al., 1998). Recent advances in high-throughput RNA sequencing (RNA-seq) have substantially increased our ability to detect RS events. Whole-cell ribosomal RNA-depleted total RNA-seq and nascent RNA-seq have been used to capture RS splicing intermediates and nascent transcripts to identify RS events(Duff et al., 2015; Hayashi et al., 2018; Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2015; Zhang et al., 2018). Nevertheless, the efficiency of capturing RS splicing

intermediates and nascent transcripts using these methods is relatively low. Therefore, new methods improving the capture efficiency are needed to study RS events.

The RS depends on a sequence motif of juxtaposed 3' and 5' splice site (AGGT). This motif is critical for RS because mutating the motif sequence disrupts the RS process and interferes with the function of the host genes (Joseph et al., 2018). However, it remains unclear how the RS splicing machinery selectively and precisely utilizes a specific AGGT site while ignoring hundreds of other AGGT sites in the same intron. Meanwhile, little is known about why some long introns use the RS mechanism for splicing, while other long introns do not. Furthermore, RS utilizes the splicing mechanism of exon definition(Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2015; Zhang et al., 2018), which requires an RS exon downstream of the AGGT site. Although RS exons are typically spliced from mature transcripts, multiple lines of evidence indicate that some RS exons can be retained in mature mRNAs(Joseph et al., 2018; Sibley et al., 2015). A group of annotated exons use RS-like mechanism for splicing(Joseph et al., 2018; Sibley et al., 2015), indicating a possible link between RS exons and annotated exons. Still, it remains unknown whether RS exons may evolve into annotated exons.

Here, we focused on nuclear RNAs to investigate RS events because RS splicing intermediates and nascent transcripts are mainly localized in the nucleus. We analyzed nuclear total RNA-seq data that we generated from the male mouse cerebral cortex, male cortical excitatory neurons, male cortical inhibitory neurons, and female cortical excitatory neurons(Johnson et al., 2017; Zhao et al., 2018). We identified novel RS sites, examined

the cell type and sex specificity of RS, and characterized the features of RS sites in mice. We also uncovered exon-like RS, which is a novel type of RS events. With a series of analysis of the sequence context of RS sites, non-RS AGGT sites, and exon-like RS, we discovered that the primary sequence context distinguishes RS sites from non-RS AGGT sites and determines the choice between RS and exon-like RS. Our findings reveal mechanistic insights into the splicing and evolution of long genes and provide a new avenue to understand the human diseases associated with long genes.

## Results

**Nuclear total RNA is enriched for nascent transcripts and RS intermediates.** To investigate whether nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates, we compared the nuclear total RNA-seq data and the whole-cell total RNA-seq data we recently generated from the cerebral cortex of 6-week-old mice(Johnson et al., 2017). We aligned the raw sequencing reads to the mouse mm10 genome and obtained 610 million uniquely mapped reads (Figure S1). Given that the total RNA-seq reads in introns indicate nascent transcripts(Ameur et al., 2011; Tilgner et al., 2012), we calculated the proportions of reads mapped to introns. We found that 77% of the uniquely mapped reads from the nuclear total RNA-seq data were localized in introns (Figure 1A and Figure S1), which was significantly higher than the 41% from the whole-cell total RNA-seq data ($P < 2.2^{-16}$, one-tailed Fisher's Exact Test). In contrast, only 23% of the uniquely mapped reads were localized in introns from poly(A) enriched messenger RNA-seq (mRNA-seq) data from the mouse cerebral cortex(Pervouchine et al., 2015) (Figure

S1). Thus, nuclear total RNA-seq captures more nascent transcripts than whole-cell total RNA-seq does.

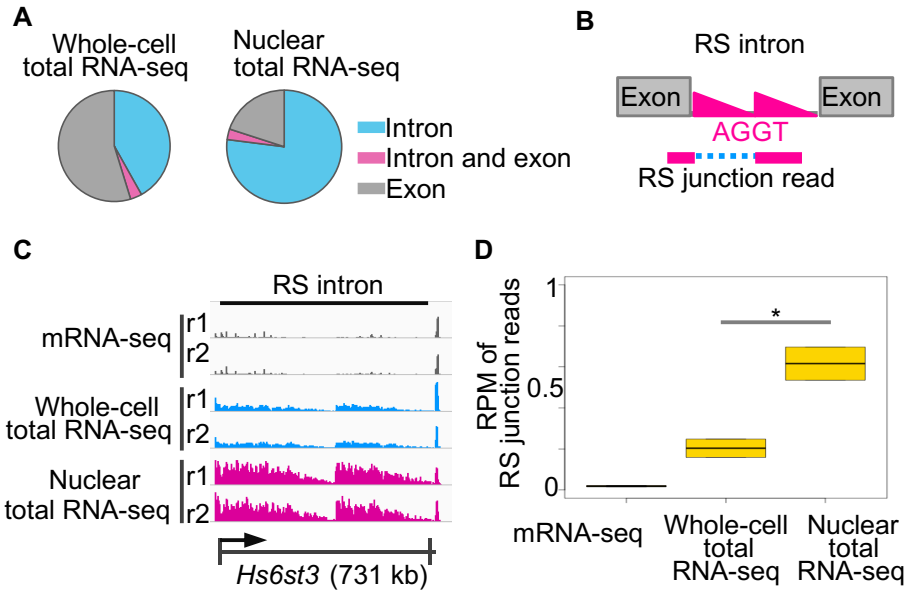


**Figure 1. Nuclear total RNA is enriched for nascent transcripts and recursive splicing (RS) intermediates.** (**A**) Pie charts of loci of uniquely mapped reads in gene regions. (**B**) Schematic of the two features of RS - the saw-tooth pattern (red triangles) and the RS junction read. (**C**) Sequencing profile at *Hs6st3* locus. r1, replicate 1. kb, kilobases. (**D**) 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.

Two features are critical for the identification of RS sites from total RNA-seq data (Figure 1B). One feature is the saw-tooth pattern of decreasing read density in the host intron (RS intron)(Duff et al., 2015; Sibley et al., 2015) (Figure 1B, red triangles), which indicates nascent transcripts and splicing status(Ameur et al., 2011; Tilgner et al., 2012). The other feature is the junction reads spanning upstream exon and the AGGT RS site (RS junction reads, Figure 1B), which indicate the RS splicing intermediates. To examine the efficiency of the two total RNA-seq methods, we compared the two RS features between the two methods in a known RS intron, the intron of *Hs6st3(Sibley et al., 2015)*. A more distinct saw-tooth pattern was observed in nuclear total RNA-seq data than that in whole-cell total RNA-seq data (Figure 1C). Furthermore, the number of RS junction

reads at *Hs6st3* was 3-fold higher in nuclear total RNA-seq data than in whole-cell total RNA-seq data (Figure 1D). Together, these results demonstrate that nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates.

**Identification of RS sites from nuclear total RNA-seq data.** To identify new RS sites, we developed a pipeline based on the RS junction reads and the saw-tooth patterns in RS introns from total RNA-seq data (Figure 2A and Figure S2). We extracted all junction reads spanning AGGT sites from the alignment results, focused on AGGT sites located in long introns (length ≥ 50 kb), selected sites enriched for RS junction reads compared to mRNA-seq, and chose sites containing 10 or more RS junction reads as RS site candidates (Figure S2). The candidates were further refined based on saw-tooth patterns in their host introns (Figure S2).

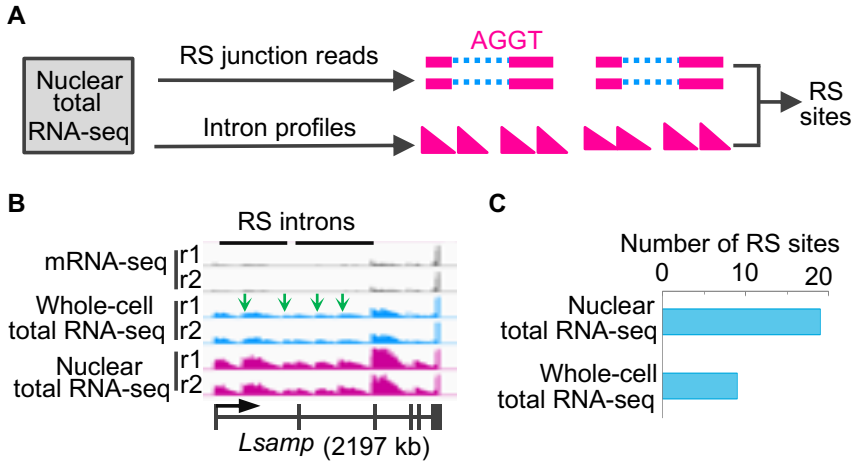


**Figure 2. Identification of RS sites from nuclear total RNA-seq data.** (**A**) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS sites. (**B**) Sequencing profile at Lsamp locus. Green arrows indicate the four novel RS sites. (**C**) Bar plot of RS sites identified utilizing the two sequencing methods in the mouse cortex.

By applying this pipeline to our nuclear total RNA-seq data, we identified 19 RS sites, which include all of the ten known RS sites in mice(Sibley et al., 2015) (Figure S3), indicating the high sensitivity of our method. Notably, 47% of RS sites we identified are

novel, including the four sites in the introns of *Lsamp* (Figure 2B, green arrows). In addition, we identified 111% more RS sites using nuclear total RNA-seq data than using whole-cell total RNA-seq data (Figure 2C and Figure S3). Together, these results demonstrate that utilizing nuclear total RNA-seq data is efficient to identify RS sites.

**Cell type and sex specificity of RS in the mouse cerebral cortex.** To investigate the cell type specificity of RS, we analyzed RS in the two major types of neurons in the mouse cerebral cortex, the excitatory neurons and the inhibitory neurons, which account for 85% and 15% of the neurons in the mouse cerebral cortex(Johnson et al., 2017). In our approach, the excitatory neurons were NeuN positive and *NeuroD6/NEX-Cre* positive, while the inhibitory neurons were NeuN positive and *NeuroD6/NEX-Cre* negative (Johnson et al., 2017). Nuclear total RNA-seq data of the two neuronal cell types were generated from the cerebral cortex of 6-week-old mice by our previous study(Johnson et al., 2017), and each cell type contained four biological replicates. Alignment of the raw data to the mouse genome resulted in 288 million uniquely mapped reads for the data of excitatory neurons and 281 million uniquely mapped reads for the data of inhibitory neurons (Figure S1). Then, we applied our pipeline to these data and identified 17 RS sites in excitatory neurons and 18 RS sites in inhibitory neurons (Figure 3A and Figure S3). Notably, all but one of the RS sites are common in both neuronal cell types (Figure 3A). The only RS site unique to inhibitory neurons resides in the *Kcnip1* gene, which is only expressed in the inhibitory neurons (Figure 3B and 3C). Therefore, these results

strongly indicate that RS is largely constitutive between these two types of neurons in the mouse cerebral cortex.

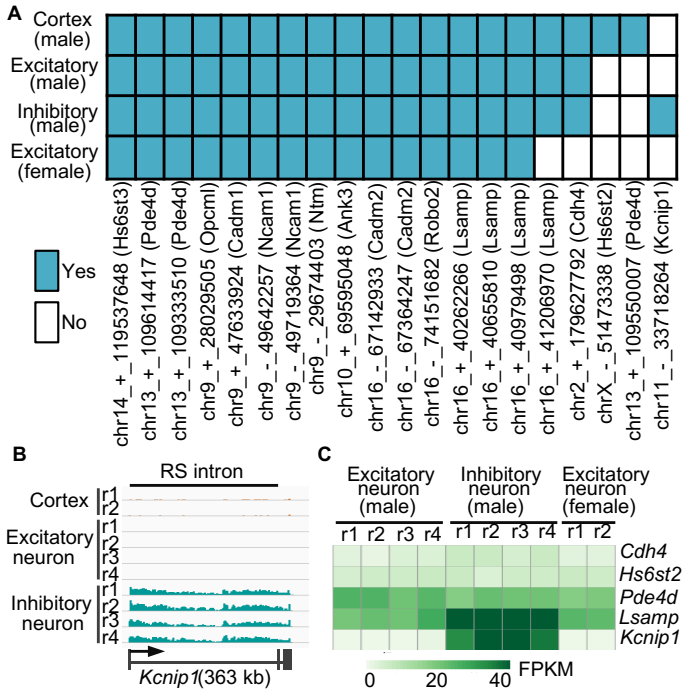


**Figure 3. Cell type and sex specificity of RS in the mouse cortex.** (**A**) Heatmap of RS sites identified in male cortex, male cortical excitatory and inhibitory neurons, and female cortical excitatory neurons. (**B**) Nuclear total RNA-seq profile (male) at Kcnip1 locus. (**C**) Heatmap of expression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

To investigate the sex specificity of RS, we also analyzed the nuclear total RNA-seq data of excitatory neurons from the cerebral cortex of female mice(Johnson et al., 2017) and identified 15 RS sites (Figure S1 and S3). All of the 15 sites are included in the 17 RS sites we identified from male excitatory neurons (Figure 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 (Figure S3), 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 the RS sites in mice.** We next investigated the characteristics of the 20 RS sites we identified in mice (Figure S3). Given the high conservation of RS mechanism among species(Duff et al., 2015; Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2015; Zhang et al., 2018), we first investigated the sequence conservation of the 600 nt regions surrounding the RS sites. In agreement with previous studies(Duff et al.,

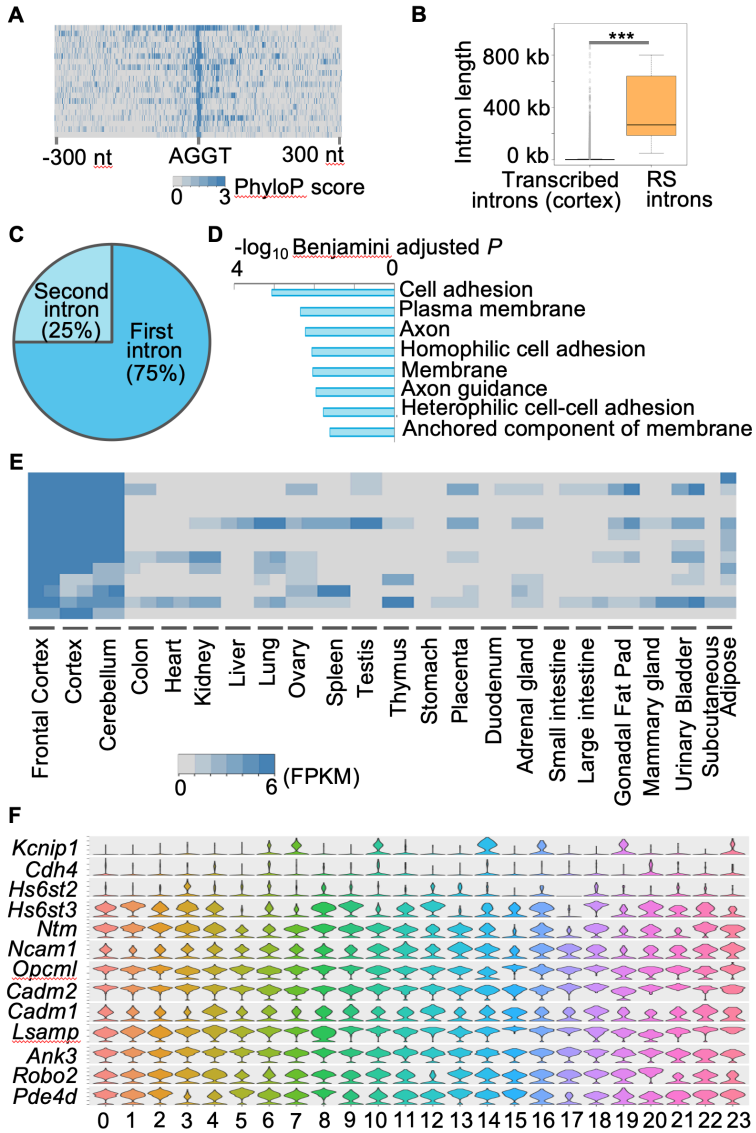


**Figure 4. Characteristics of RS in mice**. (**A**) Heatmap of phyloP score of RS sites and the flanking regions. (**B**) Boxplot of lengths of RS introns and introns transcribed in the mouse cortex. ***, P < 0.0001, one-tailed t-test. (**C**) Pie chart of locations of RS introns in host genes. (**D**) Enriched Gene Ontology terms of RS genes. (**E**) Heatmap of expression levels of RS genes in 22 mouse tissues. (**F**) Violin plot of expression levels of RS genes in the 24 cell types in the mouse cerebral cortex.

2015; Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2015), the AGGT motif at the RS sites is highly conserved across 60 vertebrate genomes, showing high phylogenetic p-value scores (phyloP scores) compared to upstream and downstream regions (Figure 4A). The RS introns, showing a median length of 267 kb and a mean length of 382 kb, were significantly longer than introns transcribed in the mouse cerebral cortex (Figure 4B). Based on the profile of histone H3 lysine 4 trimethylation in the mouse cerebral cortex(Zhao et al., 2018), we found that 75% of RS introns were the first intron, and the rest 25% were the second intron in their host genes (Figure 4C). Given the impact of first introns and long introns on transcription(Fong and Zhou, 2001; Pai et al., 2017; Shepard et al., 2009; Singh and Padgett, 2009; Swinburne et al., 2008), RS introns may also contribute to transcriptional regulation.

**RS genes are specifically expressed in the brain.** RS genes are highly expressed long genes, showing a median length of 628 kb (Figure S4A). To investigate the biological functions of RS genes, we performed Gene Ontology enrichment analysis and found that RS genes were enriched in cell adhesion, axon, membrane, and axon guidance (Figure 4D). This result indicates that RS genes are important for neuronal function and brain development. To determine whether RS genes are specifically expressed in the brain, we investigated the expression patterns of RS genes among 22 mouse tissues from the ENCODE project(Pervouchine et al., 2015). We found that RS genes are specifically expressed in brain regions including the cortex, frontal cortex, and cerebellum, but rarely in other tissues (Figure 4E). To further explore the expression patterns of RS genes in different cell types in the brain, we examined their expression by analyzing single nucleus

RNA-seq data generated from the mouse cerebral cortex(Hu et al., 2017), which profiled gene expression in 24 cell types including excitatory neurons, inhibitory neurons, astrocytes, and oligodendrocytes (Figure S4B). We found that although 77% of RS genes were constitutively expressed among most of the 24 cell types, 23% of RS genes were expressed restrictedly in specific cell types (Figure 4F). For example, *Kcnip1* is only expressed in the five subtypes of inhibitory neurons (Figure 4F and Figure S4B).

**Primary sequence context determines RS sites in RS introns.** It remains unclear how RS splicing machinery precisely utilizes a specific AGGT site while ignoring other AGGT sites in the same intron. For example, although there are 2641 AGGT sites in the RS intron of *Hs6st3*, only one specific site is always utilized by the RS splicing machinery (Figure 5A).

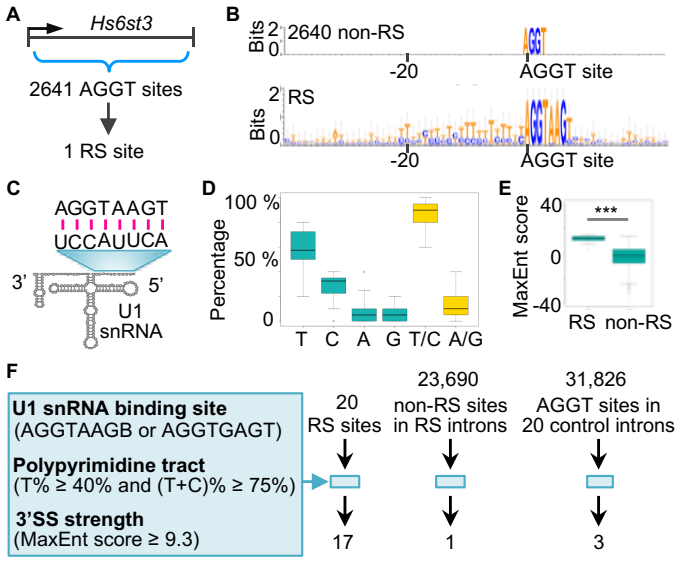


**Figure 5. Primary sequence context determines RS sites in RS introns.** (**A**) Schematic of AGGT sites in Hs6st3 intron. (**B**) Sequence logos of the 64 nt regions surrounding the 2640 non-RS AGGT sites in Hs6st3 intron and the 20 RS sites. (**C**) Schematic of the sequence base pairing between the AGGTAAGT motif and U1 snRNA. (**D**) Boxplot of the percentages of nucleotides in the 20 nt region upstream of the AGGT RS sites. (**E**) Boxplot of MaxEnt 3' splice site (3'SS) scores of RS sites and the 2640 non-RS AGGT sites. ***, P = 1.42-14, one-tailed t-test. (**F**) The criteria used to predict RS sites and the results. B represents T, G, or C.

To investigate the specificity of RS sites, we systematically compared the 2640 non-RS AGGT sites in *Hs6st3* intron with the 20 RS sites we identified. First, we examined the sequence features surrounding the AGGT sites using WebLogo(Crooks et al., 2004) and found two features specific to RS sites (Figure 5B). One feature is the enrichment of AGGTAAGT motif (Figure 5B), which complements with the 5' conserved sequence of U1 snRNA (Figure 5C). Notably, the base pairing between U1 snRNA and splice site is critical for splicing recognition(Seraphin et al., 1988). The other feature is the enrichment of thymine and cytosine in the 20 nt regions upstream of AGGT sites (Figure 5B), known as the polypyrimidine tract(Fu et al., 1988). To quantify this enrichment, we calculated the nucleotide composition of these regions and found the high percentages of thymine and cytosine in RS sites (Figure 5D and Figure S5A), but not in non-RS AGGT sites (Figure S5B). Lastly, given that RS sites are used as the 3' splice sites (3'SS) during splicing, we quantified the strengths of 3'SS using MaxEntScan(Yeo and Burge, 2004) and found significantly higher 3'SS scores of RS sites than that of non-RS AGGT sites (Figure 5E). Together, these results demonstrate that RS sites exhibit specific primary sequence features, including the U1 snRNA binding site, the polypyrimidine tract, and the strong 3'SS.

Next, we investigated whether these primary sequence features could be used to predict RS sites. We developed a pipeline to predict RS sites based on the presence of AGGT motifs and three additional criteria (Figure 5F). First, the AGGT site should have a U1 snRNA binding site, AGGTAAGB (B is T, G, or C) or AGGTGAGT (Figure 5F). Second, the 20 nt region upstream of the AGGT site should have a polypyrimidine tract. The percentage of thymine in this region should be ≥ 40%, and the combined percentage

of thymine and cytosine in this region should be ≥ 75% (Figure 5F). Third, the AGGT site should have a strong 3'SS. The MaxEnt score of 3'SS should be ≥ 9.3 (Figure 5F). The cutoff values of these criteria were obtained from the values of *Hs6st3* (Figure S5A). By applying this pipeline, we found that 85% (17) of RS sites passed the three criteria (Figure 5F and Figure S5A). In contrast, all but one of the 23,690 (99.996%) non-RS AGGT sites in the RS introns failed these criteria (Figure 5F and Figure S5C), including the 2640 sites in *Hs6st3* intron. Therefore, our prediction pipeline has an accuracy of 85% and a false discovery rate of 0.004%. Together, these data suggest that primary sequence context determines RS sites in RS introns.

**Lack of the required primary sequence context restrains RS in other long introns.** It remains unclear why some long introns utilize the RS splicing mechanism, while others do not. We hypothesized that the absence of RS mechanism in other long introns is due to the lack of the required primary sequence features at AGGT sites. To test this hypothesis, we examined the AGGT sites in these long introns. We focused on 20 control long introns, which showed the similar intron length and expression levels of host genes compared to RS introns (Figure S5D and S5E). There are 31,826 AGGT sites in the 20 control introns. Notably, all but three of them (99.99%) failed our prediction criteria of RS sites (Figure 5F and Figure S5C), suggesting that the lack of the required primary sequence context restrains RS in these long introns.

**Identification of exon-like RS events.** Recursive splicing utilizes the splicing mechanism of exon definition, which requires an RS exon following the AGGT

site(Georgomanolis et al., 2016; Hollander et al., 2016; Joseph et al., 2018; Robberson et al., 1990; Sibley et al., 2015). Given that a group of annotated exons utilize a RS-like mechanism(Joseph et al., 2018; Sibley et al., 2015), we hypothesized that RS exons might evolve into annotated exons. To test this hypothesis, we developed a pipeline to identify exon-like RS events (hereafter termed exonicRS) (Figure S6A and Figure 6A), which could be the intermediate stage of RS exons evolving into annotated exons. This pipeline is based on the assumption that exonicRS contains RS exon-like features, junction reads spanning the upstream exon (Up junction reads), and junction reads spanning the downstream exon (Down junction reads) (Figure 6A). By applying this pipeline to our nuclear total RNA-seq data from the mouse cerebral cortex, we discovered 22 exonicRS in the introns of 21 long genes (Figure S6B).

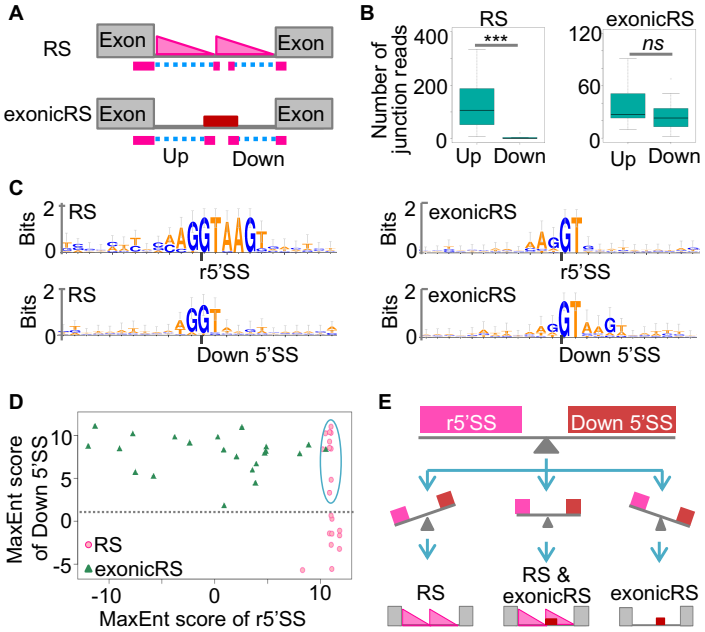


**Figure 6. Strengths of 5' splice sites (5'SS) classify RS events into RS sites and exonicRS.** (**A**) Schematic of RS sites (RS), exon-like RS events (exonicRS), the junction reads spanning upstream exon (Up), and the junction reads spanning downstream exon (Down). (**B**) Boxplots of the numbers of the Up and Down junction reads of RS and exonicRS. ***, P < 0.0001, one-tailed t-test. ns, P = 0.29, one-tailed t-test. (**C**) Sequence logos of the reconstituted 5'SS (r5'SS) and the downstream 5'SS (Down 5'SS) of RS and exonicRS. (**D**) Scatterplot of the MaxEnt scores of r5'SS and Down 5'SS of RS and exonicRS. (**E**) Model for the classification of RS events based on the strengths of r5'SS and Down 5'SS.

To illuminate the features of exonicRS, we first compared the numbers of junction reads between RS sites and exonicRS. We found that RS sites exhibited 37-fold more Up junction reads than Down junction reads (Figure 6B). In contrast, exonicRS exhibited comparable numbers of Up and Down junction reads (Figure 6B). Next, we investigated the 5' splice sites (5'SS) at the two ends of RS exons and exonicRS, the reconstituted 5'SS (r5'SS) after the first step of splicing and the downstream 5'SS (Down 5'SS). These were investigated because the competition between them may affect the inclusion of the RS exons(Joseph et al., 2018; Sibley et al., 2015; Zhang et al., 2018). We examined the sequence features of r5'SS and Down 5'SS using WebLogo(Crooks et al., 2004) and found that RS sites showed an enrichment of 5'SS motif (AGGTAAGT) at the r5'SS but not at the Down 5'SS, while exonicRS showed the opposite trend (Figure 6C). Quantification of the strengths of r5'SS and Down 5'SS using MaxEntScan(Yeo and Burge, 2004) demonstrated that the MaxEnt scores of r5'SS were significantly higher than that of Down 5'SS in RS sites (Figure S6C), while the MaxEnt scores of r5'SS were significantly lower than that of Down 5'SS in exonicRS (Figure S6C). Thus, RS sites and exonicRS exhibit distinct molecular features, suggesting that their underlying mechanisms might be different.

**Strengths of 5'SS are able to classify RS events into RS sites and exonicRS.** Given the differences of 5'SS strengths of RS sites and exonicRS, we next asked whether the strengths of r5'SS and Down 5'SS were able to distinguish the two types of RS events. We plotted the MaxEnt scores of 5'SS of the 42 RS events we identified, including the 20 RS sites and the 22 exonicRS. RS events were classified into three categories based on

the scores of r5'SS and Down 5'SS (Figure 6D and Figure S6B). One category is exonicRS, which exhibits high Down 5'SS scores and low r5'SS scores (Figure 6D, green triangles). Another category is RS sites, which exhibits high r5'SS scores and low Down 5'SS scores (Figure 6D, pink dots). There is also a third category that contains ten RS sites and one exonicRS (Figure 6D, blue oval region). Further analyses revealed that these ten RS sites in the third category all contained Down junction reads (Figure S6B), suggesting that the third category is a combination of RS sites and exonicRS (Figure 6E). This combination is further supported by the observation that the exonicRS in the third category exhibits a weak saw-tooth pattern, which failed to pass our stringent criteria in our initial identification of RS sites (Figure S6D). Notably, both of the two RS sites (*Cadm1* and *Ank3*), which have been experimentally confirmed to utilize the exon definition mechanism(Sibley et al., 2015), were classified into the third category (Figure S6B). Together, these results support a mechanism that the strengths of 5'SS are able to classify RS events into RS sites and exonicRS (Figure 6E).

## Discussion

The human genome contains more than 4000 genes that are longer than 100 kb. Since gene transcription relies on RNA polymerase II moving along the entire gene region, a long gene consumes more time and resources to make a transcript than a short gene does. To overcome the length constraint, long genes may utilize specific mechanisms to regulate their transcription and co-transcriptional processes, such as splicing. RS is a splicing mechanism specific to long introns in long genes. In this study, we developed an efficient pipeline to investigate RS, identified new RS sites, and

discovered a novel type of RS events. Our discovery of exonicRS indicates that long genes may acquire novel exons via the RS exons, which is in consistent with the observation that long introns exhibit a high rate of creating new exons during evolution(Roy et al., 2008). Given that more than 6000 human exons are putative RS exons(Blazquez et al., 2018), RS may represent a commonly used mechanism for long gene splicing, particularly for the splicing of alternative exons and putative RS exons flanked by long introns.

Each of the RS introns contains hundreds of AGGT sites. Given that the RS splicing machinery utilizes a specific AGGT site while ignoring others(Duff et al., 2015; Hayashi et al., 2018; Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2016; Sibley et al., 2015; Zhang et al., 2018), the RS sites may carry unique features compared to the non-RS AGGT sites. We found that RS sites exhibit specific primary sequence context of U1 snRNA binding site and polypyrimidine tract, which may recruit RNA binding proteins (RBPs), such as exon junction complex and polypyrimidine tract binding proteins(Blazquez et al., 2018; Cai et al., 2020; Duff et al., 2015; Patton et al., 2020; Ule and Blencowe, 2019; Van Nostrand et al., 2020; Xue et al., 2009), to facilitate the recognition of 3'SS. Thus, profiling RBP bindings and RNA structures around the RS and non-RS AGGT sites are necessary to further illustrate the molecular basis of RS.

RS exons are typically skipped from mature transcripts. The findings that some annotated exons utilizes the RS-like splicing mechanism(Blazquez et al., 2018; Sibley et al., 2015) suggest a possible link between RS exons and annotated exons. We

discovered a novel type of RS events termed exonicRS. The exonicRS exhibits features similar to both RS exons and annotated exons, indicating that RS exons may evolve into annotated exons, and that exonicRS may be the intermediate stage between the two. This may explain the observation that the numbers of RS sites in *Drosophila melanogaster* are about 15 times more than that in humans(Duff et al., 2015; Joseph et al., 2018; Pai et al., 2018; Sibley et al., 2015; Zhang et al., 2018), and the numbers of RS-like annotated exons in *Drosophila* are 2~100 times less than that in humans(Blazquez et al., 2018; Joseph et al., 2018; Sibley et al., 2015). Therefore, further investigation is required to dissect the relationship among RS exons, exonicRS, and novel exons in evolutionarily distinct species.

RS genes are specifically expressed in the brain and 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(Ferreira et al., 2008; Iqbal et al., 2013; Sanders et al., 2012; Schulze et al., 2009). Also, *NTM* encodes neurotrimin and is linked to autism spectrum disorders and attention deficit hyperactivity disorder(Brevik et al., 2016; Maruani et al., 2015). The *PDE4D* encodes phosphodiesterase 4D and is linked to schizophrenia, psychosis, acrodysostosis, and neuroticism(Lee et al., 2012; Shifman et al., 2008; Sinha et al., 2019). Notably, PDE4D Inhibitors are in clinical trials for the treatment of Alzheimer's disease and Fragile X syndrome(Gurney et al., 2019; Pan et al., 2019). Given that the disruption of the RS process interfered with the RS gene function and caused abnormality in the central nervous system(Joseph et al., 2018), further investigation will be necessary to

illuminate whether RS mechanism contributes to the pathophysiology of these human brain disorders.

**Acknowledgements**

We thank Dr. Raddy Ramos, Dr. Weikang Cai, Dr. Lars Udo-Bellner, and members of the Long Gene Lab 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**

J.Y.Z. conceived the project, designed the experiments, performed the computational analyses, and wrote the manuscript.

**Declaration of Interests**

The authors declare no competing interests.

**Figure Legends**

**Figure 1** Nuclear total RNA is enriched for nascent transcripts and recursive splicing (RS) intermediates. **(A)** Pie charts of loci of uniquely mapped reads in gene regions. **(B)** Schematic of the two features of RS - the saw-tooth pattern (red triangles) and the RS junction read. **(C)** Sequencing profile at *Hs6st3* locus. r1, replicate 1. kb, kilobases. **(D)** 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.

**Figure 2** Identification of RS sites from nuclear total RNA-seq data. **(A)** Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS sites. **(B)** Sequencing profile at *Lsamp* locus. Green arrows indicate the four novel RS sites. **(C)** Bar plot of RS sites identified utilizing the two sequencing methods in the mouse cortex.

**Figure 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.

**Figure 4** Characteristics of RS in mice. **(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)** Enriched Gene Ontology terms of RS genes. **(E)** Heatmap of expression levels of RS genes in 22 mouse tissues. **(F)** Violin plot of expression levels of RS genes in the 24 cell types in the mouse cerebral cortex.

**Figure 5** Primary sequence context determines RS sites in RS introns. **(A)** Schematic of AGGT sites in *Hs6st3* intron. **(B)** Sequence logos of the 64 nt regions surrounding the 2640 non-RS AGGT sites in *Hs6st3* intron and the 20 RS sites. **(C)** Schematic of the sequence base pairing between the AGGTAAGT motif and U1 snRNA. **(D)** Boxplot of the percentages of nucleotides in the 20 nt region upstream of the AGGT RS sites. **(E)** Boxplot of MaxEnt 3' splice site (3'SS) scores of RS sites and the 2640 non-RS AGGT sites. ***, $P = 1.42^{-14}$, one-tailed t-test. **(F)** The criteria used to predict RS sites and the results. B represents T, G, or C.

**Figure 6** Strengths of 5' splice sites (5'SS) classify RS events into RS sites and exonicRS. **(A)** Schematic of RS sites (RS), exon-like RS events (exonicRS), the junction reads spanning upstream exon (Up), and the junction reads spanning downstream exon (Down). **(B)** Boxplots of the numbers of the Up and Down junction reads of RS and exonicRS. ***, $P < 0.0001$, one-tailed t-test. *ns*, $P = 0.29$, one-tailed t-test. **(C)** Sequence logos of the reconstituted 5'SS (r5'SS) and the downstream 5'SS (Down 5'SS) of RS and exonicRS. **(D)** Scatterplot of the MaxEnt scores of r5'SS and Down 5'SS of RS and exonicRS. **(E)** Model for the classification of RS events based on the strengths of r5'SS and Down 5'SS.

## STAR Methods

**Nuclear total RNA-seq data analysis.** Raw data in sra files were downloaded from the EBI European Nucleotide Archive database(Amid et al., 2020) using the accession numbers listed in Figure S1. The fastq-dump.2.9.6 of NCBI SRA ToolKit was used to extract the FASTQ files using the parameter of "--split-3". STAR(Dobin and Gingeras, 2015) 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(Li et al., 2009) 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(Ramirez et al., 2016) 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(Robinson et al., 2011).

**Genome annotation.** The gtf file of mouse genome annotation was downloaded from the Ensembl release 93(Yates et al., 2020).

**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 Figure S2. Briefly, AGGT sites located in introns longer than or equal to 50 kb were extracted. 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 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.

**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 Ontology enrichment analysis.** Gene Ontology enrichment analysis was performed by the Functional Annotation Tool of the DAVID Bioinformatics Resources 6.8(Huang da et al., 2009). Genes with a mean exonic FPKM ≥ 1 in the nuclear total RNA-seq data were used as the background genes. The Benjamini corrected *P* values were used to obtain the significantly enriched Gene Ontology terms.

**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).

**Control long introns.** Introns transcribed in the mouse cerebral cortex were sorted by intron lengths. Only the highly expressed introns (host gene FPKM > 20) were retained, and the top 20 longest introns were used as the control long introns.

**Single nucleus RNA-seq data.** The expression levels of RS genes in the 24 cell types in the mouse cerebral cortex were obtained from the LongGeneDB database (https://longgenedb.com).

**WebLogo analysis.** WebLogo 3(Crooks et al., 2004) (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)(Yeo and Burge, 2004). 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)(Yeo and Burge, 2004). 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 Figure S6B.

**Pipeline to identify exonicRS.** The schematic of this pipeline is shown in Figure S6A. 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 exonicRS candidates. The exonicRS candidates were further refined as exonicRS if the host intron showed an exon-like but not saw-tooth like pattern.

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

**Availability of data and materials**

Sequencing data that support the findings of this study are available in NCBI GEO database with the accession codes listed in Figure S1. The custom code supporting the conclusions of this article is available in the GitHub repository, https://github.com/Jerry-Zhao/RS2020.

**Supplemental Information**

Supplemental Information can be found online.

**References**

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol *18*, 1435-1440.

Amid, C., Alako, B.T.F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P.W., Holt, S., Hussein, A., Ivanov, E.*, et al.* (2020). The European Nucleotide Archive in 2019. Nucleic Acids Res *48*, D70-D76.

Blazquez, L., Emmett, W., Faraway, R., Pineda, J.M.B., Bajew, S., Gohr, A., Haberman, N., Sibley, C.R., Bradley, R.K., Irimia, M.*, et al.* (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. Mol Cell *72*, 496-509 e499.

Brevik, E.J., van Donkelaar, M.M., Weber, H., Sanchez-Mora, C., Jacob, C., Rivero, O., Kittel-Schneider, S., Garcia-Martinez, I., Aebi, M., van Hulzen, K.*, et al.* (2016). Genome-wide analyses of aggressiveness in attention-deficit hyperactivity disorder. Am J Med Genet B Neuropsychiatr Genet *171*, 733-747.

Burnette, J.M., Miyamoto-Sato, E., Schaub, M.A., Conklin, J., and Lopez, A.J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics *170*, 661-674.

Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X.*, et al.* (2020). RIC-seq for global in situ profiling of RNA-RNA spatial interactions. Nature.

Cook-Andersen, H., and Wilkinson, M.F. (2015). Molecular biology: Splicing does the two-step. Nature *521*, 300-301.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics *51*, 11 14 11-11 14 19.

Duff, M.O., Olson, S., Wei, X., Garrett, S.C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S.E., and Graveley, B.R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature *521*, 376-379.

Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419-437.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., Fan, J., Kirov, G., Perlis, R.H., Green, E.K.*, et al.* (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet *40*, 1056-1058.

Fong, Y.W., and Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. Nature *414*, 929-933.

Fu, X.Y., Ge, H., and Manley, J.L. (1988). The role of the polypyrimidine stretch at the SV40 early pre-mRNA 3' splice site in alternative splicing. EMBO J *7*, 809-817.

Georgomanolis, T., Sofiadis, K., and Papantonis, A. (2016). Cutting a Long Intron Short: Recursive Splicing and Its Implications. Front Physiol *7*, 598.

Gurney, M.E., Nugent, R.A., Mo, X., Sindac, J.A., Hagen, T.J., Fox, D., 3rd, O'Donnell, J.M., Zhang, C., Xu, Y., Zhang, H.T.*, et al.* (2019). Design and Synthesis of Selective Phosphodiesterase 4D (PDE4D) Allosteric Inhibitors for the Treatment of Fragile X Syndrome and Other Brain Disorders. J Med Chem *62*, 4884-4901.

Hatton, A.R., Subramaniam, V., and Lopez, A.J. (1998). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell *2*, 787-796.

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun *9*, 619.

Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A.R., and Ast, G. (2016). How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? Trends Genet *32*, 596-606.

Hu, P., Fabyanic, E., Kwon, D.Y., Tang, S., Zhou, Z., and Wu, H. (2017). Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq. Mol Cell *68*, 1006-1015 e1007.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Iqbal, Z., Vandeweyer, G., van der Voet, M., Waryah, A.M., Zahoor, M.Y., Besseling, J.A., Roca, L.T., Vulto-van Silfhout, A.T., Nijhof, B., Kramer, J.M.*, et al.* (2013). Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. Hum Mol Genet *22*, 1960-1970.

Johnson, B.S., Zhao, Y.T., Fasolino, M., Lamonica, J.M., Kim, Y.J., Georgakilas, G., Wood, K.H., Bu, D., Cui, Y., Goffin, D.*, et al.* (2017). Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome. Nat Med *23*, 1203-1214.

Joseph, B., Kondo, S., and Lai, E.C. (2018). Short cryptic exons mediate recursive splicing in Drosophila. Nat Struct Mol Biol *25*, 365-371.

Kelly, S., Georgomanolis, T., Zirkel, A., Diermeier, S., O'Reilly, D., Murphy, S., Langst, G., Cook, P.R., and Papantonis, A. (2015). Splicing of many human genes involves sites embedded within introns. Nucleic Acids Res *43*, 4721-4732.

Lee, H., Graham, J.M., Jr., Rimoin, D.L., Lachman, R.S., Krejci, P., Tompson, S.W., Nelson, S.F., Krakow, D., and Cohn, D.H. (2012). Exome sequencing identifies PDE4D mutations in acrodysostosis. Am J Hum Genet *90*, 746-751.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Maruani, A., Huguet, G., Beggiato, A., ElMaleh, M., Toro, R., Leblond, C.S., Mathieu, A., Amsellem, F., Lemiere, N., Verloes, A.*, et al.* (2015). 11q24.2-25 micro-rearrangements in autism spectrum disorders: Relation to brain structures. Am J Med Genet A *167A*, 3019-3030.

Pai, A.A., Henriques, T., McCue, K., Burkholder, A., Adelman, K., and Burge, C.B. (2017). The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. Elife *6*.

Pai, A.A., Paggi, J.M., Yan, P., Adelman, K., and Burge, C.B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. PLoS Genet *14*, e1007588.

Pan, T., Xie, S., Zhou, Y., Hu, J., Luo, H., Li, X., and Huang, L. (2019). 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 *29*, 2150-2152.

Patton, R.D., Sanjeev, M., Woodward, L.A., Mabin, J.W., Bundschuh, R., and Singh, G. (2020). Chemical crosslinking enhances RNA immunoprecipitation for efficient identification of binding sites of proteins that photo-crosslink poorly with RNA. RNA.

Pervouchine, D.D., Djebali, S., Breschi, A., Davis, C.A., Barja, P.P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L.H.*, et al.* (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun *6*, 5903.

Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet *6*, e1001236.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol *10*, 84-94.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Roy, M., Kim, N., Xing, Y., and Lee, C. (2008). The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA *14*, 2261-2273.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L.*, et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.

Schulze, T.G., Detera-Wadleigh, S.D., Akula, N., Gupta, A., Kassem, L., Steele, J., Pearl, J., Strohmaier, J., Breuer, R., Schwarz, M.*, et al.* (2009). Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. Mol Psychiatry *14*, 487-491.

Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nat Rev Genet *17*, 19-32.

Seraphin, B., Kretzner, L., and Rosbash, M. (1988). A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. EMBO J *7*, 2533-2538.

Shepard, S., McCreary, M., and Fedorov, A. (2009). The peculiarities of large intron splicing in animals. PLoS One *4*, e7853.

Shifman, S., Bhomra, A., Smiley, S., Wray, N.R., James, M.R., Martin, N.G., Hettema, J.M., An, S.S., Neale, M.C., van den Oord, E.J.*, et al.* (2008). A whole genome association study of neuroticism using DNA pooling. Mol Psychiatry *13*, 302-312.

Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nat Rev Genet *17*, 407-421.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J.*, et al.* (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371-375.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol *16*, 1128-1133.

Sinha, V., Ukkola-Vuoti, L., Ortega-Alonso, A., Torniainen-Holm, M., Therman, S., Tuulio-Henriksson, A., Jylha, P., Kaprio, J., Hovatta, I., Isometsa, E.*, et al.* (2019). Variants in regulatory elements of PDE4D associate with major mental illness in the Finnish population. Mol Psychiatry.

Swinburne, I.A., Miguez, D.G., Landgraf, D., and Silver, P.A. (2008). Intron length increases oscillatory periods of gene expression in animal cells. Genes Dev *22*, 2342-2346.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res *22*, 1616-1625.

Ule, J., and Blencowe, B.J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Mol Cell *76*, 329-345.

Van Nostrand, E.L., Pratt, G.A., Yee, B.A., Wheeler, E.C., Blue, S.M., Mueller, J., Park, S.S., Garcia, K.E., Gelboin-Burkhart, C., Nguyen, T.B.*, et al.* (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol *21*, 90.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H.*, et al.* (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell *36*, 996-1006.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R.*, et al.* (2020). Ensembl 2020. Nucleic Acids Res *48*, D682-D688.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

Zhang, X.O., Fu, Y., Mou, H., Xue, W., and Weng, Z. (2018). The temporal landscape of recursive splicing during Pol II transcription elongation in human cells. PLoS Genet *14*, e1007579.

Zhao, Y.T., Kwon, D.Y., Johnson, B.S., Fasolino, M., Lamonica, J.M., Kim, Y.J., Zhao, B.S., He, C., Vahedi, G., Kim, T.H.*, et al.* (2018). Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains. Genome Res *28*, 933-942.