Nuclear total RNA sequencing reveals primary sequence context of recursive splicing in long genes

Sohyun Moon[1], Jacob Vazquez[1], Jerry Yingtao Zhao[1*]

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

*Correspondence: yzhao47@nyit.edu

Abstract

Background: 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.

Results: We identified novel RS sites and discovered that RS is constitutive between excitatory and inhibitory neurons and between sexes in the mouse cerebral cortex. We found that the primary sequence context, including the U1 snRNA binding site, the polypyrimidine tract, and a strong 3' splice site, distinguishes the RS AGGT site from hundreds of non-RS AGGT sites in the same intron. Moreover, we uncovered a new type of exon-like RS events termed exonicRS.

Conclusions: We demonstrate that nuclear total RNA sequencing is an efficient approach to identify RS events. We find the importance of the primary sequence context in the definition of RS AGGT sites. The exonicRS may represent an intermediate stage of RS sites evolving into annotated exons. Overall, our findings provide novel insights into the mechanisms of RS in long genes.

Keywords

Recursive splicing, RS, nuclear total RNA-seq, total RNA-seq, RS exon, long intron, long gene.

Background

Recursive splicing (RS) is a splicing mechanism that is specific to long introns in long genes[1-10]. RS removes 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. Whole-cell ribosomal RNA-depleted total RNA sequencing (total RNA-seq) and nascent RNA-seq have been used to capture RS splicing intermediates[3, 4, 7-10], but the efficiency is relatively low. Therefore, new approaches are needed to identify RS events.

RS depends on a sequence motif of juxtaposed 3' and 5' splice site (AGGT). However, each RS intron contains hundreds to thousands of AGGT sites. It remains unclear how the RS splicing machinery selectively and precisely utilizes a specific AGGT site while ignoring other AGGT sites in the same intron. In addition, a group of annotated exons (RS exons) use RS-like mechanism for splicing[4, 7], suggesting a possible link between RS and RS exons, but the details of this link remain elusive.

Here, we focused on nuclear RNAs to investigate RS events because RS splicing intermediates and nascent transcripts are mainly localized in the nucleus. We reanalyzed the nuclear total RNA-seq data that we generated from the mouse cerebral cortex[11, 12]. 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 AGGT sites, non-RS AGGT sites, and exon-like RS, we discovered that the primary sequence

context distinguishes RS AGGT sites from non-RS AGGT sites and determines the choice between RS and exon-like RS. Our findings provide novel insights into the mechanism and evolution of RS in long genes.

Results

Nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates. Two features are critical for the identification of RS sites from total RNA-seq data (Fig. 1a). One feature is the saw-tooth pattern of decreasing read density in the host intron (RS intron)[3, 4] (Fig. 1a, red triangles), which indicates nascent transcripts and splicing status[13, 14]. The other feature is the junction reads spanning upstream exon and the RS AGGT site (RS junction reads, Fig. 1a), which indicate the RS splicing intermediates. Given that nascent transcripts and RS splicing intermediates are localized in the nucleus, we asked whether nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates compared to whole-cell total RNA (Fig. 1b). To answer this question, we reanalyzed the whole-cell total RNA-seq and nuclear total RNA-seq data (Figure S1a in Additional file 1) that we generated from the cerebral cortex of 6-week-old mice[11, 12].

We first examined nascent transcripts in whole-cell total RNA-seq data, nuclear total RNA-seq data, and poly(A) enriched messenger RNA-seq (mRNA-seq) data from the mouse cerebral cortex[15]. We calculated the proportions of reads mapped to introns because total RNA-seq reads in introns indicate nascent transcripts[13, 14]. We found that 77% of the uniquely mapped reads from the nuclear total RNA-seq data were localized in introns (Figure S1a,b in Additional file 1), 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 mRNA-seq data (Figure S1a in Additional file 1). Thus, nuclear total RNA-seq captures more nascent transcripts than whole-cell total RNA-seq does.

We next compared the two RS features (Fig. 1a) between the two total RNA-seq methods in a known RS intron, the intron of *Hs6st3[4]*. A more distinct saw-tooth pattern was observed in nuclear total RNA-seq data than that in whole-cell total RNA-seq data (Fig. 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 (Fig. 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 (Fig. 1e and Figure S1c in Additional file 1). 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 S1c in Additional file 1). The candidates were further refined based on saw-tooth patterns in their host introns (Figure S1c in Additional file 1).

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[4] (Figure S1d in Additional file 1), indicating the high sensitivity of our method. Notably, 47% of RS sites we identified are

116 novel, including the four sites in the introns of *Lsamp* (Fig. 1f, green arrows). In addition,
117 we identified 111% more RS sites using nuclear total RNA-seq data than using whole-
118 cell total RNA-seq data (Fig. 1g and Figure S1d in Additional file 1). Together, these
119 results demonstrate that nuclear total RNA-seq is an efficient approach to identify RS
120 sites.
121
122 **Cell type specificity of RS in the mouse cerebral cortex.** To investigate the cell type
123 specificity of RS, we analyzed RS in the two major types of neurons in the mouse cerebral
124 cortex, the excitatory neurons and the inhibitory neurons, which account for 85% and 15%
125 of the neurons in the mouse cerebral cortex[11]. We reanalyzed the nuclear total RNA-
126 seq data we generated from the two neuronal cell types (Figure S1a in Additional file 1).
127 We applied our pipeline to these data and identified 17 RS sites in excitatory neurons and
128 18 RS sites in inhibitory neurons (Fig. 2a and Figure S1d in Additional file 1). Notably, all
129 but one of the RS sites are common in both neuronal cell types (Fig. 2a). The only RS
130 site unique to inhibitory neurons resides in the *Kcnip1* gene, which is only expressed in
131 the inhibitory neurons (Fig. 2b and 2c). Therefore, these results indicate that RS is largely
132 constitutive between these two types of neurons in the mouse cerebral cortex.
133
134 **Sex specificity of RS in the mouse cerebral cortex.** To investigate the sex specificity
135 of RS, we also reanalyzed the nuclear total RNA-seq data of excitatory neurons from the
136 cerebral cortex of female mice[11] and identified 15 RS sites (Figure S1d in Additional file
137 1). All of the 15 sites are included in the 17 RS sites we identified from male excitatory
138 neurons (Fig. 2a). 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 S1d in Additional file 1), 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 S1d in Additional file 1). Given the high conservation of RS mechanism among species[3, 4, 7-9], we first investigated the sequence conservation of the 600 nt regions surrounding the RS sites. In agreement with previous studies[3, 4, 7, 8], 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 (Fig. 3a). 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 (Fig. 3b). Based on the profile of histone H3 lysine 4 trimethylation in the mouse cerebral cortex[12], we found that 75% of RS introns were the first intron, and the rest 25% were the second intron in their host genes (Fig. 3c).

RS genes are specifically expressed in the brain. RS genes are highly expressed long genes, showing a median length of 628 kb (Figure S1e in Additional file 1). 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[15]. We found that RS genes are specifically expressed in brain regions including the cortex, frontal cortex, and cerebellum, but rarely in other tissues (Fig. 3d). To further explore the

163 expression patterns of RS genes in different cell types in the brain, we examined their
164 expression by analyzing single nucleus RNA-seq data generated from the mouse cerebral
165 cortex[16], which profiled gene expression in 24 cell types including excitatory neurons,
166 inhibitory neurons, astrocytes, and oligodendrocytes (Figure S1f in Additional file 1). We
167 found that although 77% of RS genes were constitutively expressed among most of the
168 24 cell types, 23% of RS genes were expressed restrictedly in specific cell types (Fig.
169 3e). For example, *Kcnip1* is only expressed in the five subtypes of inhibitory neurons (Fig.
170 3e and Figure S1f in Additional file 1).
171
172 **The primary sequence features are different between RS and non-RS AGGT sites**
173 **in the same introns.** It remains unclear how RS splicing machinery precisely utilizes a
174 specific AGGT site but ignoring other AGGT sites in the same intron. For example,
175 although there are 2641 AGGT sites in the intron of *Hs6st3*, only one AGGT site is used
176 by the RS splicing machinery (Fig. 4a).
177 To identify the difference between RS and non-RS AGGT sites, we systematically
178 compared the 2640 non-RS AGGT sites in *Hs6st3* intron with all the 20 RS sites we
179 identified. First, we examined the sequence features surrounding the AGGT sites using
180 WebLogo[17] and found two features specific to RS AGGT sites (Fig. 4b). One feature is
181 the enrichment of AGGTAAGT motif (Fig. 4b), which complements with the 5' conserved
182 sequence of U1 snRNA (Fig. 4c). Notably, the base pairing between U1 snRNA and splice
183 site is critical for splicing recognition[18]. The other feature is the enrichment of thymine
184 and cytosine in the 20 nt regions upstream of RS AGGT sites (Fig. 4b), known as the
185 polypyrimidine tract[19]. To quantify this enrichment, we calculated the nucleotide

composition of these regions and found the high percentages of thymine and cytosine in RS AGGT sites, but not in non-RS AGGT sites (Fig. 4d). Lastly, given that RS AGGT sites are used as the 3' splice sites (3'SS) during splicing, we quantified the strengths of 3'SS using MaxEntScan[20] and found significantly higher 3'SS scores of RS AGGT sites than that of non-RS AGGT sites (Fig. 4e). Together, these results demonstrate that RS AGGT sites exhibit specific primary sequence features compared to non-RS AGGT sites.

Computational prediction of RS sites based on primary sequences alone. Next, we investigated whether primary sequence features alone could be used to predict RS sites. We developed a computational pipeline to predict RS sites based on the presence of AGGT motifs and three additional sequence features (Fig. 4f, blue box). First, the AGGT site should have a U1 snRNA binding site, AGGTAAGB (B is T, G, or C) or AGGTGAGT (Fig. 4f). 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% (Fig. 4f). Third, the AGGT site should have a strong 3'SS. The MaxEnt score of 3'SS should be ≥ 9.3 (Fig. 4f). The cutoff values of these criteria were obtained from the values of *Hs6st3* (Figure S2a in Additional file 2).

We applied the computational prediction pipeline on the 23,710 AGGT sites in the 20 RS introns. Eighteen AGGTs sites passed the criteria (Fig. 4f), which include 85% (17 of 20) of RS AGGT sites that we identified from the sequencing data (Fig. 4f and Figure S2a in Additional file 2). In contrast, 99.996% (all but one of the 23,690) of non-RS AGGT sites in RS introns failed these criteria (Fig. 4f and Figure S2b in Additional file 2). Thus, our

209 computational prediction pipeline has an accuracy of 85% and a false discovery rate of
210 0.004%. Together, these data suggest that RS AGGT sites are largely determined by
211 their primary sequence context.

212 It remains unclear that all long introns contain AGGT sites, but most of long introns
213 do not use the RS splicing mechanism. We speculated that this is because AGGT sites
214 in non-RS long introns lack the required primary sequence context of RS. To test this, we
215 applied our computational prediction pipeline to 20 non-RS introns, which showed the
216 similar intron length and expression levels of host genes compared to the 20 RS introns
217 we identified (Figure S2c,2d in Additional file 2). There are 31,826 AGGT sites in the 20
218 non-RS introns (Fig. 4f). Notably, 99.99% (all but three) of these AGGT sites failed our
219 computational prediction criteria of RS sites (Fig. 4f and Figure S2b in Additional file 2),
220 indicating that the lack of the required primary sequence context restrains RS in these
221 long introns.

222

223 **Identification of exon-like RS events.** A group of annotated exons (RS exons) utilize a
224 RS-like mechanism for splicing[4, 7], suggesting a possible link between RS and
225 annotated RS exons. Given that RS uses the splicing mechanism of exon definition[4, 7,
226 21-23], we hypothesized that RS may evolve into annotated RS exons through an
227 intermediate stage of exon-like RS events (hereafter termed exonicRS) (Fig. 5a). To test
228 this hypothesis, we developed a pipeline to identify exonicRS (Figure S3a in Additional
229 file 3). This pipeline is based on the assumption that exonicRS contains RS exon-like
230 features, junction reads spanning the upstream exon (Up junction reads), junction reads
231 spanning the downstream exon (Down junction reads), and no saw-tooth pattern (Fig.

5a). By applying this pipeline to our nuclear total RNA-seq data, we discovered 22 exonicRS in the introns of 21 long genes (Figure S3b in Additional file 3).

To illuminate the features of exonicRS, we first analyzed the numbers of junction reads in RS and exonicRS. We found that RS exhibited 37-fold more Up junction reads than Down junction reads (Fig. 5b). In contrast, exonicRS exhibited comparable numbers of Up and Down junction reads (Fig. 5b). Next, we investigated the 5' splice sites (5'SS) at the two ends of the exons at RS sites 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 be associated with the inclusion of the RS exons in mature transcripts[4, 7, 9]. We examined the sequence features of r5'SS and Down 5'SS using WebLogo[17] and found that RS 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 (Fig. 5c). Quantification of the strengths of r5'SS and Down 5'SS using MaxEntScan[20] demonstrated that the MaxEnt scores of r5'SS were significantly higher than that of Down 5'SS in RS sites (Fig. 5d), while the MaxEnt scores of r5'SS were significantly lower than that of Down 5'SS in exonicRS (Fig. 5d). Taken together, these results demonstrate that RS and exonicRS exhibit distinct molecular features.

Strengths of 5'SS are able to distinguish RS and exonicRS. We next asked whether the strengths of r5'SS and Down 5'SS were able to distinguish RS and exonicRS. We plotted the MaxEnt scores of 5'SS of the 20 RS and the 22 exonicRS (Fig. 5e). Unexpectedly, they were classified into three categories based on the 5'SS scores (Figure 5e and Figure S3b in Additional file 3). One category is exonicRS, which exhibits high

Down 5'SS scores and low r5'SS scores (Fig. 5e, green triangles). Another category is RS, which exhibits high r5'SS scores and low Down 5'SS scores (Fig. 5e, pink dots). There is also a third category that contains ten RS and one exonicRS (Fig. 5e, blue oval region). Further analyses revealed that these ten RS sites in the third category all contained Down junction reads (Figure S3b in Additional file 3), suggesting that the third category is a combination of RS and exonicRS (Fig. 5a). 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 S3c in Additional file 3). Notably, both of the two RS sites (*Cadm1* and *Ank3*), which have been experimentally confirmed to utilize the exon definition mechanism in mammals[4], were classified into the third category (Figure S3b in Additional file 3). Together, these results support that the strengths of 5'SS are able to distinguish RS and exonicRS (Fig. 5f).

Discussion

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 use 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 identify RS from nuclear total RNA-seq data, investigated the primary sequence context of RS, and discovered a novel type of RS events. Our identification of 20 RS sites from high-depth nuclear total RNA-seq data suggests that RS is likely a special splicing mechanism only for a small portion of introns in the mammalian genome.

RS introns are mainly (75%) first introns. Given that first introns, particularly their lengths, are critical for the transcriptional activities of host genes[24-28], RS mechanism may also contribute to the transcriptional regulation.

Each long intron contains hundreds to thousands of AGGT sites, suggesting that AGGT motif alone could not determine the RS mechanism. Previous studies reported that RS AGGT sites exhibit specific sequence context, but it remains unclear whether the sequence context alone could predict RS. We systematically examined the sequence context of RS sites and developed a computational pipeline to predict RS with high accuracy and low false discovery rate. Our findings indicate that RS is largely determined by the primary sequence context. Notably, several RS sites failed our prediction pipeline, suggesting that other factors, such as RNA binding proteins and RNA structures[3, 29-34], may also play a role in the definition of RS mechanism. Thus, further investigations are necessary to integrate the primary sequence context, RNA binding protein binding, and RNA structures to illustrate the molecular basis of RS.

Long introns exhibit a high rate of creating new exons during evolution[35], but the underlying mechanism remains unclear. Our discovery of exonicRS indicates that long introns may acquire novel exons via the RS mechanism. This is supported by the findings that more than 6000 human annotated exons are putative RS exons[29]. In addition, the numbers of RS sites in *Drosophila melanogaster* are about 15 times more than that in humans[3, 4, 7-9], but the numbers of RS-like annotated exons in *Drosophila* are 2~100 times less than that in humans[4, 7, 29]. These observations further support our indication of RS sites evolving into annotated exons. Therefore, future studies are required to

investigate RS, exonicRS, and RS exons in evolutionarily distinct species using approaches such as nuclear total RNA-seq.

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[36-39]. Also, *NTM* encodes neurotrimin and is linked to autism spectrum disorders and attention deficit hyperactivity disorder[40, 41]. The *PDE4D* encodes phosphodiesterase 4D and is linked to schizophrenia, psychosis, acrodysostosis, and neuroticism[42-44]. Notably, PDE4D Inhibitors are in clinical trials for the treatment of Alzheimer's disease and Fragile X syndrome[45, 46]. Given that the disruption of the RS process interfered with the RS gene function and caused abnormality in the central nervous system[7], further investigation will be necessary to illuminate whether RS mechanism contributes to the pathophysiology of these human brain disorders.

Conclusions

In this study, we reveal the molecular mechanism of RS in long introns of long genes. Our results highlighted that nuclear total RNA-seq is an efficient approach to investigate RS. We develop a novel pipeline to identify RS events, characterize the cell type specificity and genomic features of RS sites, and discover a novel type of RS events. Through analysis of primary sequence, we demonstrate that RS is largely determined by the primary sequence context, thus providing novel insights into the RS mechanism. Our discovery of exonicRS indicates a new mechanism by which long genes could acquire new exons. Overall, our findings provide mechanistic insights into the splicing and

323 evolution of long genes and reveal a new avenue to understand the human diseases
324 associated with these long genes.
325
326 **Methods**
327 **Statistical analysis.** All statistical analyses were performed in the R software version
328 3.6.1 (https://www.r-project.org/).
329 **Nuclear total RNA-seq data analysis.** Raw data in sra files were downloaded from the
330 EBI European Nucleotide Archive database[47] using the accession numbers listed in
331 Figure S1. The fastq-dump.2.9.6 of NCBI SRA ToolKit was used to extract the FASTQ
332 files using the parameter of "--split-3". STAR[48] was used to map the FASTQ raw reads
333 into mouse mm10 genome using the parameters of "--runThreadN 40 --
334 outFilterMultimapNmax 1 --outFilterMismatchNmax 3". The samtools view[49] was used
335 to convert sam files into bam files. The samtools sort was used to sort the bam files. The
336 samtools index was used to index the sorted bam files. The bamCoverage[50] was used
337 to convert the sorted bam files into strand-specific bigwig files. The bamCoverage
338 parameters that were used included "--filterRNAstrand forward --binSize 1 -p 40 -o" for
339 plus strand and "--filterRNAstrand reverse --binSize 1 -p 40 -o" for minus strand.
340 **Sequencing data visualization.** All sequencing data, including the bigwig files and bam
341 files, were visualized in the IGV_2.8.2 genome browser[51].
342 **Genome annotation.** The gtf file of mouse genome annotation was downloaded from the
343 Ensembl release 93[52].
344 **Junction reads.** A read pair is considered as a junction read if its CIGAR in sam files
345 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 S1c in Additional file 1. 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 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[17] (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)[20]. 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)[20]. 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 S3b in Additional file 3.

Pipeline to identify exonicRS. The schematic of this pipeline is shown in Figure S3a in Additional file 3. 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.

Abbreviations:

RS: recursive splicing; mRNA: messenger RNA; kb: kilobase; RNA-seq: RNA sequencing; mRNA-seq: poly(A) enriched messenger RNA sequencing; phyloP: phylogenetic p-value; 3'SS: 3' splice site; 5'SS; 5' splice site; RPM, reads per million uniquely mapped reads.

Declarations:

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the in NCBI GEO database with the accession codes listed in Figure S1a in Additional file 1. The custom code supporting the conclusions of this article is available in the GitHub repository, https://github.com/Jerry-Zhao/RS2020.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Authors' contributions

SM, JV, and JYZ curated data and wrote the manuscript. SM performed the experiments. JYZ conceived the project, designed the experiments, performed the computational analyses. All authors read and approved the final manuscript.

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.

References

1. 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):787-796.
2. 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):661-674.
3. Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, Plocik A, Celniker SE, Graveley BR: **Genome-wide identification of zero nucleotide recursive splicing in Drosophila**. *Nature* 2015, **521**(7552):376-379.
4. Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J *et al*: **Recursive splicing in long vertebrate genes**. *Nature* 2015, **521**(7552):371-375.
5. Cook-Andersen H, Wilkinson MF: **Molecular biology: Splicing does the two-step**. *Nature* 2015, **521**(7552):300-301.
6. Kelly S, Georgomanolis T, Zirkel A, Diermeier S, O'Reilly D, Murphy S, Langst G, Cook PR, Papantonis A: **Splicing of many human genes involves sites embedded within introns**. *Nucleic Acids Res* 2015, **43**(9):4721-4732.
7. Joseph B, Kondo S, Lai EC: **Short cryptic exons mediate recursive splicing in Drosophila**. *Nat Struct Mol Biol* 2018, **25**(5):365-371.
8. 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.
9. 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.
10. 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.
11. Johnson BS, Zhao YT, Fasolino M, Lamonica JM, Kim YJ, Georgakilas G, Wood KH, Bu D, Cui Y, Goffin D *et al*: **Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome**. *Nat Med* 2017, **23**(10):1203-1214.

12. Zhao YT, Kwon DY, Johnson BS, Fasolino M, Lamonica JM, Kim YJ, Zhao BS, He C, Vahedi G, Kim TH *et al*: **Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains**. *Genome Res* 2018, **28**(7):933-942.
13. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L: **Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain**. *Nat Struct Mol Biol* 2011, **18**(12):1435-1440.
14. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R: **Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs**. *Genome Res* 2012, **22**(9):1616-1625.
15. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See LH *et al*: **Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression**. *Nat Commun* 2015, **6**:5903.
16. Hu P, Fabyanic E, Kwon DY, Tang S, Zhou Z, Wu H: **Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq**. *Mol Cell* 2017, **68**(5):1006-1015 e1007.
17. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.
18. Seraphin B, Kretzner L, Rosbash M: **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* 1988, **7**(8):2533-2538.
19. Fu XY, Ge H, Manley JL: **The role of the polypyrimidine stretch at the SV40 early pre-mRNA 3' splice site in alternative splicing**. *EMBO J* 1988, **7**(3):809-817.
20. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals**. *J Comput Biol* 2004, **11**(2-3):377-394.
21. Robberson BL, Cote GJ, Berget SM: **Exon definition may facilitate splice site selection in RNAs with multiple exons**. *Mol Cell Biol* 1990, **10**(1):84-94.
22. Georgomanolis T, Sofiadis K, Papantonis A: **Cutting a Long Intron Short: Recursive Splicing and Its Implications**. *Front Physiol* 2016, **7**:598.
23. Hollander D, Naftelberg S, Lev-Maor G, Kornblihtt AR, Ast G: **How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?** *Trends Genet* 2016, **32**(10):596-606.
24. Pai AA, Henriques T, McCue K, Burkholder A, Adelman K, Burge CB: **The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture**. *Elife* 2017, **6**.
25. Shepard S, McCreary M, Fedorov A: **The peculiarities of large intron splicing in animals**. *PLoS One* 2009, **4**(11):e7853.
26. Singh J, Padgett RA: **Rates of in situ transcription and splicing in large human genes**. *Nat Struct Mol Biol* 2009, **16**(11):1128-1133.
27. Swinburne IA, Miguez DG, Landgraf D, Silver PA: **Intron length increases oscillatory periods of gene expression in animal cells**. *Genes Dev* 2008, **22**(17):2342-2346.
28. Fong YW, Zhou Q: **Stimulatory effect of splicing factors on transcriptional elongation**. *Nature* 2001, **414**(6866):929-933.

29. Blazquez L, Emmett W, Faraway R, Pineda JMB, Bajew S, Gohr A, Haberman N, Sibley CR, Bradley RK, Irimia M *et al*: **Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing**. *Mol Cell* 2018, **72**(3):496-509 e499.
30. Patton RD, Sanjeev M, Woodward LA, Mabin JW, Bundschuh R, Singh G: **Chemical crosslinking enhances RNA immunoprecipitation for efficient identification of binding sites of proteins that photo-crosslink poorly with RNA**. *RNA* 2020.
31. Cai Z, Cao C, Ji L, Ye R, Wang D, Xia C, Wang S, Du Z, Hu N, Yu X *et al*: **RIC-seq for global in situ profiling of RNA-RNA spatial interactions**. *Nature* 2020.
32. Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H *et al*: **Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping**. *Mol Cell* 2009, **36**(6):996-1006.
33. Ule J, Blencowe BJ: **Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution**. *Mol Cell* 2019, **76**(2):329-345.
34. Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, Park SS, Garcia KE, Gelboin-Burkhart C, Nguyen TB *et al*: **Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins**. *Genome Biol* 2020, **21**(1):90.
35. 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):2261-2273.
36. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL *et al*: **De novo mutations revealed by whole-exome sequencing are strongly associated with autism**. *Nature* 2012, **485**(7397):237-241.
37. Iqbal Z, Vandeweyer G, van der Voet M, Waryah AM, Zahoor MY, Besseling JA, Roca LT, Vulto-van Silfhout AT, Nijhof B, Kramer JM *et al*: **Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders**. *Hum Mol Genet* 2013, **22**(10):1960-1970.
38. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK *et al*: **Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder**. *Nat Genet* 2008, **40**(9):1056-1058.
39. Schulze TG, Detera-Wadleigh SD, Akula N, Gupta A, Kassem L, Steele J, Pearl J, Strohmaier J, Breuer R, Schwarz M *et al*: **Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder**. *Mol Psychiatry* 2009, **14**(5):487-491.
40. Maruani A, Huguet G, Beggiato A, ElMaleh M, Toro R, Leblond CS, Mathieu A, Amsellem F, Lemiere N, Verloes A *et al*: **11q24.2-25 micro-rearrangements in autism spectrum disorders: Relation to brain structures**. *Am J Med Genet A* 2015, **167A**(12):3019-3030.
41. Brevik EJ, van Donkelaar MM, Weber H, Sanchez-Mora C, Jacob C, Rivero O, Kittel-Schneider S, Garcia-Martinez I, Aebi M, van Hulzen K *et al*: **Genome-wide analyses of aggressiveness in attention-deficit hyperactivity disorder**. *Am J Med Genet B Neuropsychiatr Genet* 2016, **171**(5):733-747.
42. 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*: **Variants in regulatory elements of PDE4D associate with major mental illness in the Finnish population**. *Mol Psychiatry* 2019.

43. Lee H, Graham JM, Jr., Rimoin DL, Lachman RS, Krejci P, Tompson SW, Nelson SF, Krakow D, Cohn DH: **Exome sequencing identifies PDE4D mutations in acrodysostosis**. *Am J Hum Genet* 2012, **90**(4):746-751.
44. Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, Hettema JM, An SS, Neale MC, van den Oord EJ *et al*: **A whole genome association study of neuroticism using DNA pooling**. *Mol Psychiatry* 2008, **13**(3):302-312.
45. Pan T, Xie S, Zhou Y, Hu J, Luo H, Li X, Huang L: **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):2150-2152.
46. Gurney ME, Nugent RA, Mo X, Sindac JA, Hagen TJ, Fox D, 3rd, O'Donnell JM, Zhang C, Xu Y, Zhang HT *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):4884-4901.
47. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E *et al*: **The European Nucleotide Archive in 2019**. *Nucleic Acids Res* 2020, **48**(D1):D70-D76.
48. Dobin A, Gingeras TR: **Mapping RNA-seq Reads with STAR**. *Curr Protoc Bioinformatics* 2015, **51**:11 14 11-11 14 19.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
50. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis**. *Nucleic Acids Res* 2016, **44**(W1):W160-165.
51. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat Biotechnol* 2011, **29**(1):24-26.
52. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R *et al*: **Ensembl 2020**. *Nucleic Acids Res* 2020, **48**(D1):D682-D688.

Figure Legends

Fig. 1 Nuclear total RNA-seq is efficient to identify recursive splicing (RS) events. **(a)** Schematic of the two features of RS - the saw-tooth pattern (red triangles) and the RS junction read. **(b)** Schematic of the isolation of different types of RNA. **(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.

591 *, *P* = 0.03, one-tailed t-test. **(e)** Schematic of the pipeline utilizing nuclear total RNA-
592 seq data to identify RS sites. **(f)** Sequencing profile at *Lsamp* locus. Green arrows
593 indicate the four novel RS sites. **(g)** Bar plot of RS sites identified utilizing the two
594 sequencing methods in the mouse cortex.
595
596 **Fig. 2** Cell type and sex specificity of RS in the mouse cortex. **(a)** Heatmap of RS sites
597 identified in male cortex, male cortical excitatory and inhibitory neurons, and female
598 cortical excitatory neurons. **(b)** Nuclear total RNA-seq profile (male) at *Kcnip1* locus. **(c)**
599 Heatmap of expression levels of five genes in three cell types. FPKM, fragment per
600 million uniquely mapped reads per kilobase of exonic region.
601
602 **Fig. 3** Characteristics of RS in mice. **(a)** Heatmap of phyloP score of RS sites and the
603 flanking regions. **(b)** Boxplot of lengths of RS introns and introns transcribed in the
604 mouse cortex. ***, P < 0.0001, one-tailed t-test. **(c)** Pie chart of locations of RS introns
605 in host genes. **(d)** Heatmap of expression levels of RS genes in 22 mouse tissues. **(e)**
606 Violin plot of expression levels of RS genes in the 24 cell types in the mouse cerebral
607 cortex.
608
609 **Fig. 4** Primary sequence context distinguishes RS AGGT site from hundreds of non-RS
610 AGGT sites in the same intron. **(a)** Schematic of the 2641 AGGT sites in the *Hs6st3*
611 intron. Only one AGGT site is used by RS. **(b)** Sequence logos of the 64 nt regions
612 surrounding the 2640 non-RS AGGT sites in *Hs6st3* intron and the 20 RS AGGT sites.
613 **(c)** Schematic of the sequence base pairing between the AGGTAAGT motif and U1

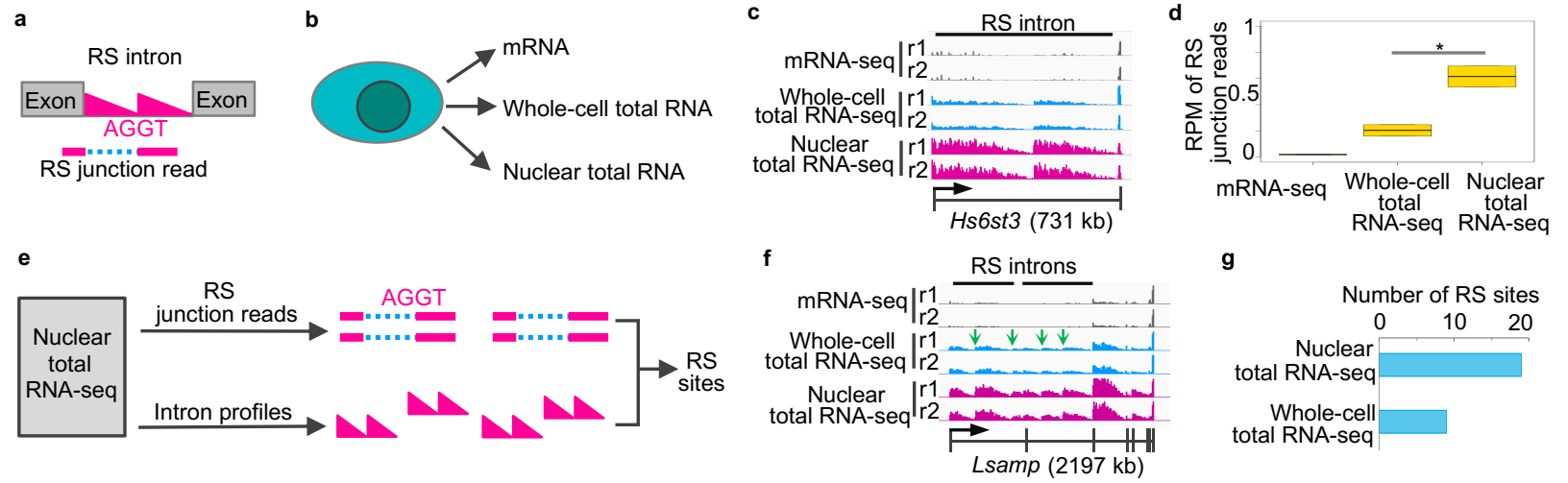
snRNA. **(d)** Boxplots of the percentages of nucleotides in the 20 nt region upstream of the 2640 non-RS AGGT sites and the 20 RS AGGT sites. **(e)** Boxplot of MaxEnt 3' splice site (3'SS) scores of the 20 RS AGGT sites and the 2640 non-RS AGGT sites. ***, $P = 1.42^{-14}$, one-tailed t-test. **(f)** A computational pipeline to predict RS from intronic AGGT sites. The criteria are listed in the blue box (left). B represents T, G, or C.

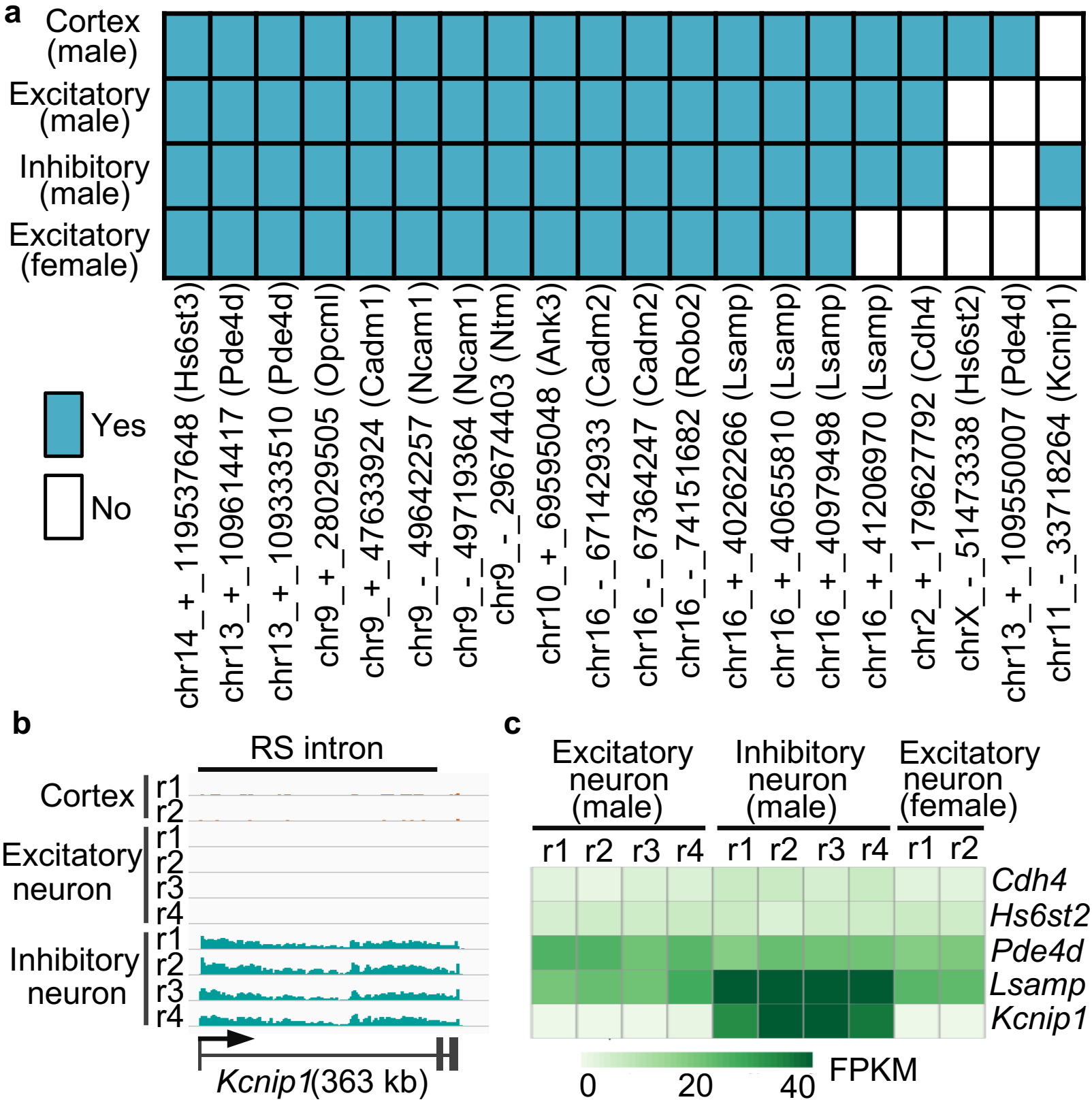
Fig. 5 The exonicRS may represent the intermediate stage of RS evolving into annotated RS exons. **(a)** Schematic of RS sites (RS) evolving into RS exons via the exon-like RS events (exonicRS). Up, the junction reads spanning upstream exon; Down, the junction reads spanning downstream exon. **(b)** 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)** Boxplot of MaxEnt scores of r5'SS and Down 5'SS of RS and exonicRS. ****, $P < 7.9^{-6}$, one-tailed t-test. **(e)** Scatterplot of the MaxEnt scores of r5'SS and Down 5'SS of RS and exonicRS. **(f)** Model for the classification of RS events based on the strengths of r5'SS and Down 5'SS.

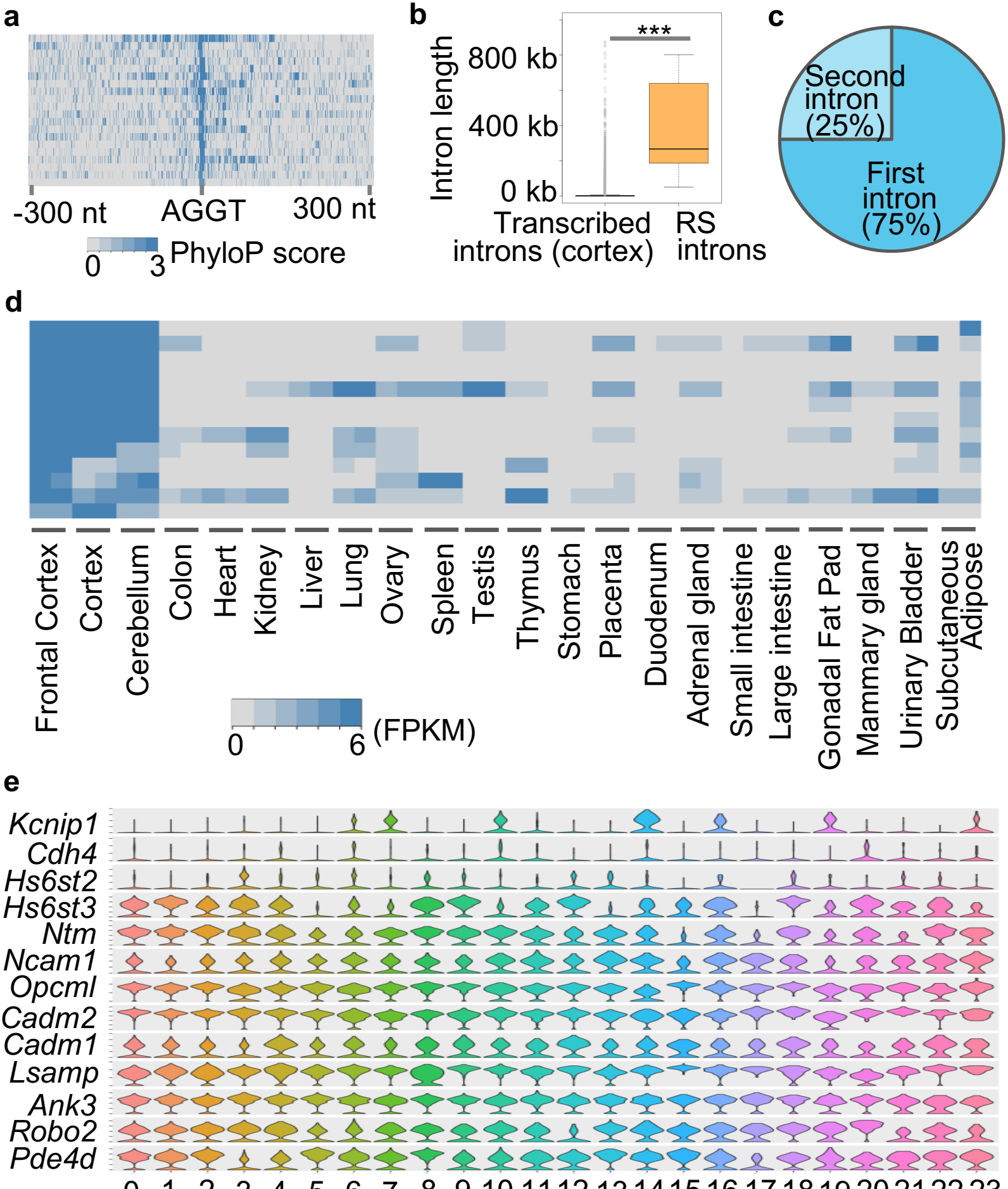
Supplemental Information

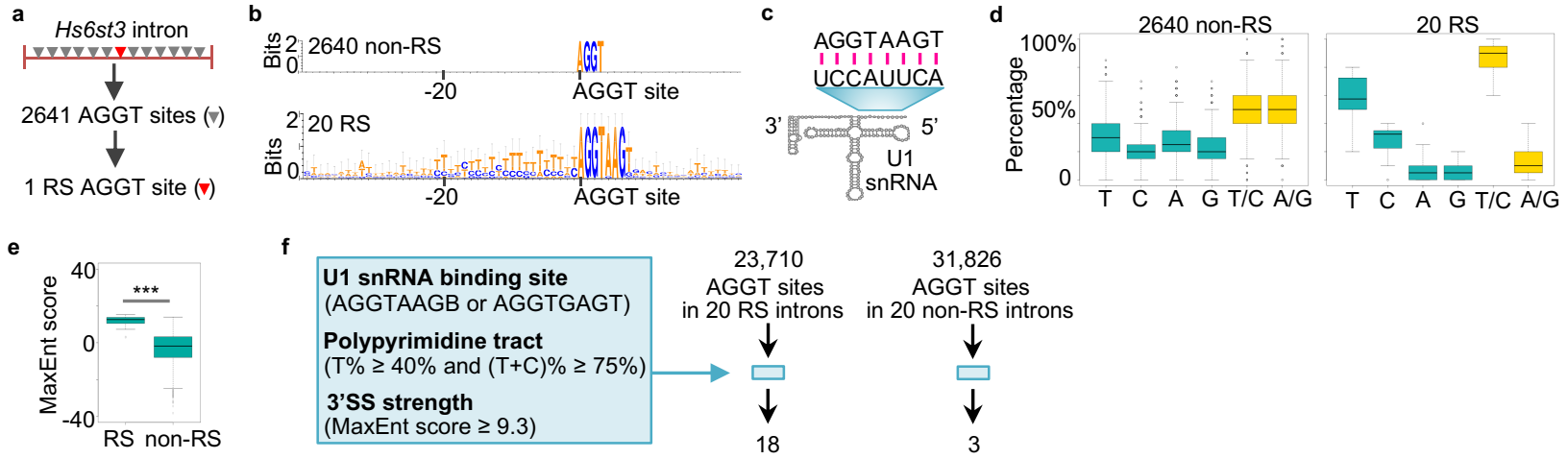
Additional file 1: Figure S1. Novel RS sites. (**a**) The mapping statistics and access numbers of RNA-seq data utilized in this study. (**b**) Pie charts of loci of uniquely mapped reads in gene regions. (**c**) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS sites. (**d**) Heatmap of numbers of RS junction reads at each RS site in

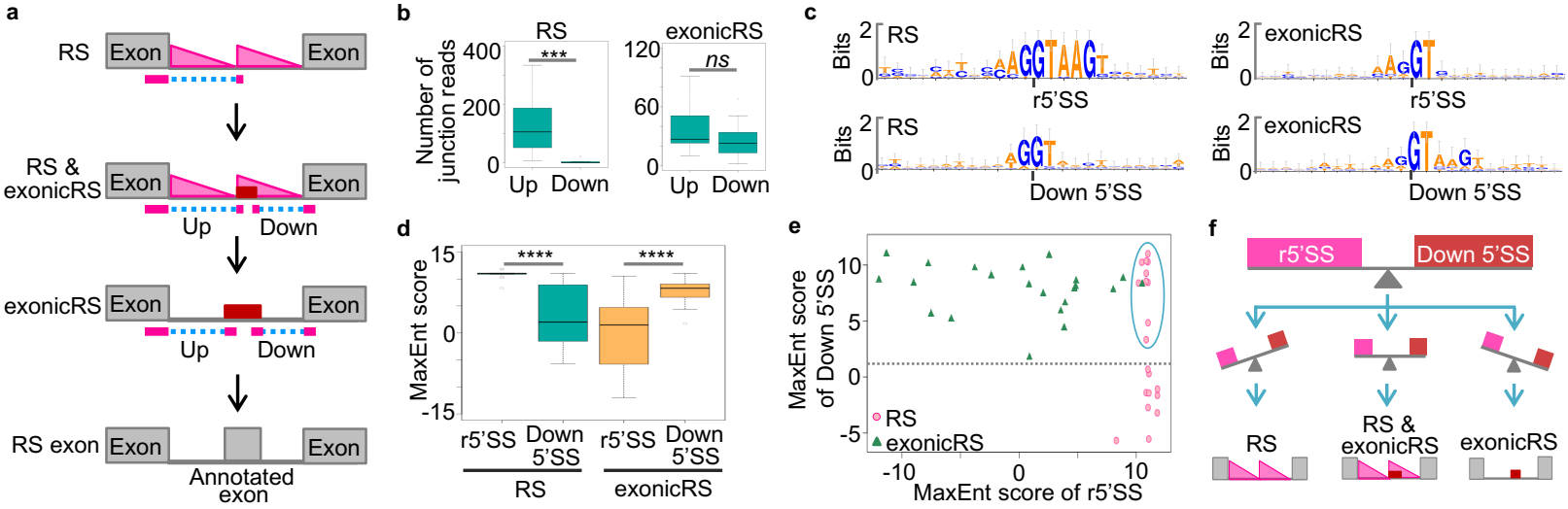
637 different total RNA-seq data sets. A green box indicates that the RS site was identified in
638 that data set. (**e**) Boxplots of the gene lengths (up) and expression levels (down) of genes
639 transcribed in the mouse cerebral cortex (transcribed gene) and RS genes. *P* indicates *P*
640 value, one-tailed t-test. (**f**) The 24 cell types in the mouse cerebral cortex revealed by
641 single nucleus RNA-seq data.
642
643 **Additional file 2: Figure S2**. Primary sequence context of RS sites. (**a**) The sequence
644 motifs, nucleotide percentages, and 3'SS MaxEnt scores of the 20 RS sites. The top 17
645 RS sites passed our prediction criteria of RS sites, while the bottom three RS sites failed.
646 (**b**) The information of the non-RS AGGTs that passed our prediction criteria of RS sites.
647 The top non-RS AGGT site resides in the RS intron of *Ank3*, while the bottom three non-
648 RS AGGT sites reside in the three control long introns. (**c**) Profiles of nuclear total RNA-
649 seq at the host genes of the 20 control long introns. Black bars indicate the control long
650 introns. (**d**) Boxplots of the intron lengths (up) and host-gene expression levels (down) of
651 introns transcribed in the mouse cerebral cortex (transcribed intron), RS introns, and
652 control long introns.
653
654 **Additional file 3: Figure S3**. Identification of exonicRS. (**a**) Schematic of the pipeline
655 utilizing nuclear total RNA-seq data to identify exonicRS. (**b**) Tables of information of
656 genomic loci, junction reads, 5'SS sequences, 5'SS MaxEnt scores, and classifications
657 of exonicRS (up) and RS sites (down). (**c**) Sequencing profile at *Magi1* locus. Green
658 arrows indicate the RS AGGT loci.
659





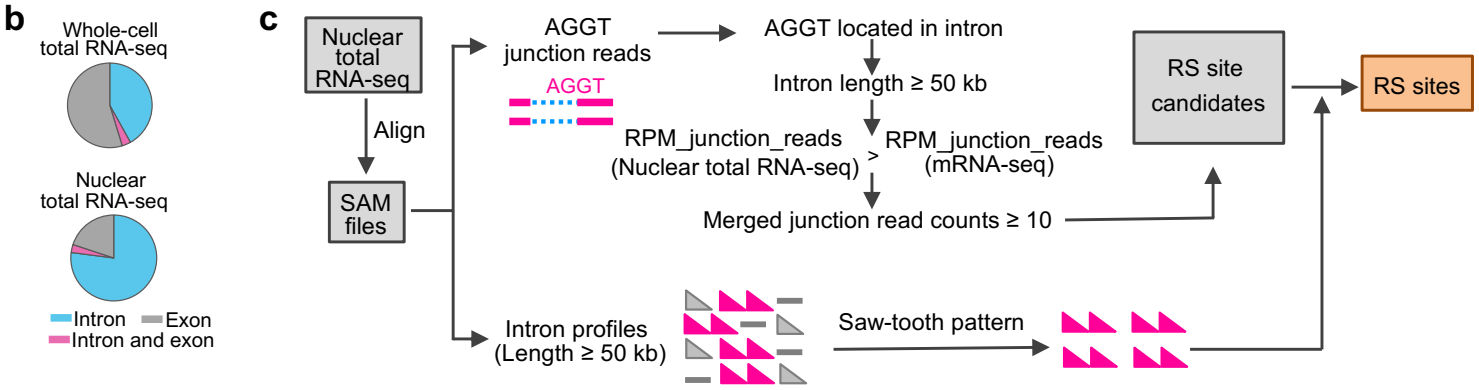






a

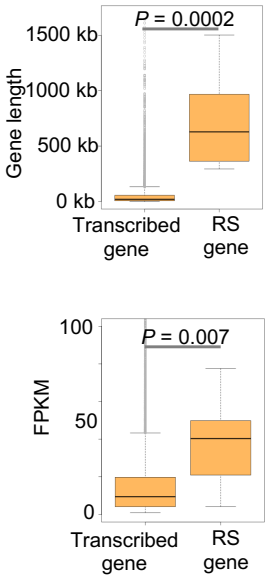
	Exon	Intron	Intron and exon	Intergenic	Totally uniquely mapped reads	NCBI SRA access number
mRNA-seq rep 1	169,199,705	52,536,837	9,942,820	18,955,236	250,634,598	SRR5048041
mRNA-seq rep 2	169,212,702	54,408,558	9,418,272	24,448,180	257,487,712	SRR5048042
Whole cell total RNA-seq rep 1	28,025,020	21,276,188	1,708,014	4,537,328	55,546,550	SRR3679862
Whole cell total RNA-seq rep 2	24,025,585	18,552,805	1,481,557	3,816,417	47,876,364	SRR3679863
Nuclear total RNA-seq rep 1	12,802,711	55,975,659	2,066,992	7,376,058	78,221,420	SRR3679866
Nuclear total RNA-seq rep 2	84,581,864	291,193,877	12,151,804	40,616,263	428,543,808	SRR9202881
Excitatory neurons rep 1	7,493,883	46,181,238	1,421,238	5,474,411	60,570,770	SRR3679830
Excitatory neurons rep 2	7,592,500	58,702,729	1,502,461	7,037,646	74,835,336	SRR3679831
Excitatory neurons rep 3	7,028,945	49,007,305	1,469,493	5,869,141	63,374,884	SRR3679832
Excitatory neurons rep 4	8,665,201	71,350,839	1,771,990	8,080,848	89,868,878	SRR3679833
Inhibitory neurons rep 1	10,620,896	51,820,218	1,846,752	5,923,736	70,211,602	SRR3679842
Inhibitory neurons rep 2	11,526,209	61,516,565	1,991,467	7,354,451	82,388,692	SRR3679843
Inhibitory neurons rep 3	10,531,020	44,429,100	1,591,223	5,326,315	61,877,658	SRR3679844
Inhibitory neurons rep 4	8,673,268	50,820,368	1,494,799	6,176,575	67,165,010	SRR3679845
Female excitatory neurons rep 1	12,995,467	56,964,384	1,920,090	7,096,375	78,976,316	SRR3679869
Female excitatory neurons rep 2	9,548,905	49,872,640	1,579,266	6,385,437	67,386,248	SRR3679870



d

Gene		Whole cell	Nuclear	Excitatory	Inhibitory	Excitatory (female)	Novel
chr9_-_29674405	Ntm	24	357	144	113	49	Yes
chr9_+_28029505	Opcml	57	283	139	89	55	-----
chr16_-_67364249	Cadm2	37	264	98	104	49	-----
chr10_+_69595048	Ank3	14	143	149	97	71	-----
chr16_-_67142935	Cadm2	31	170	56	60	43	-----
chr14_+_119537648	Hs6st3	14	171	23	25	29	-----
chr9_-_49642259	Ncam1	12	113	33	37	18	-----
chr13_+_109333510	Pde4d	6	101	28	16	20	-----
chr16_-_74151684	Robo2	4	83	25	28	13	-----
chr9_+_47633924	Cadm1	10	60	25	21	24	-----
chr9_-_49719366	Ncam1	6	72	15	22	11	Yes
chr13_+_109614417	Pde4d	5	28	23	24	15	-----
chr13_+_109550007	Pde4d	2	33	5	9	4	Yes
chr16_+_40262266	Lsamp	20	144	77	132	31	Yes
chr16_+_40979498	Lsamp	12	88	33	47	11	Yes
chr16_+_40655810	Lsamp	6	64	19	55	15	Yes
chr16_+_41206970	Lsamp	8	75	27	31	7	Yes
chr2_+_179627792	Cdh4	0	21	11	13	5	Yes
chrX_-_51473340	Hs6st2	3	20	2	1	3	Yes
chr11_-_33718266	Kcnip1	1	8	1	70	0	Yes

e



f

0: layer 2/3 excitatory neurons
1: layer 4 excitatory neurons
2: layer 5 excitatory neurons
3: striatum MSNs
4: layer 2 excitatory neurons
5: layer 6 excitatory neurons
6: GAD2+ inhibitory neurons
7: PV+ inhibitory neurons
8: layer 5 excitatory neurons
9: layer 5 excitatory neurons
10: SST+ inhibitory neurons
11: oligodendrocytes
12: layer 4 excitatory neurons
13: layer 6 excitatory neurons
14: VIP+ inhibitory neurons
15: unknown cells
16: isocortex neurons
17: hippocampus neurons
18: unknown excitatory neurons
19: NPY+ inhibitory neurons
20: thalamus neurons
21: claustrum neurons
22: astrocytes
23: oligodendrocyte precursor cells

Figure S1: Novel RS sites. (**a**) The mapping statistics and access numbers of RNA-seq data utilized in this study. (**b**) Pie charts of loci of uniquely mapped reads in gene regions. (**c**) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS 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. (**e**) Boxplots of the gene lengths (up) and expression levels (down) of genes transcribed in the mouse cerebral cortex (transcribed gene) and RS genes. *P* indicates *P* value, one-tailed t-test. (**f**) The 24 cell types in the mouse cerebral cortex revealed by single nucleus RNA-seq data.

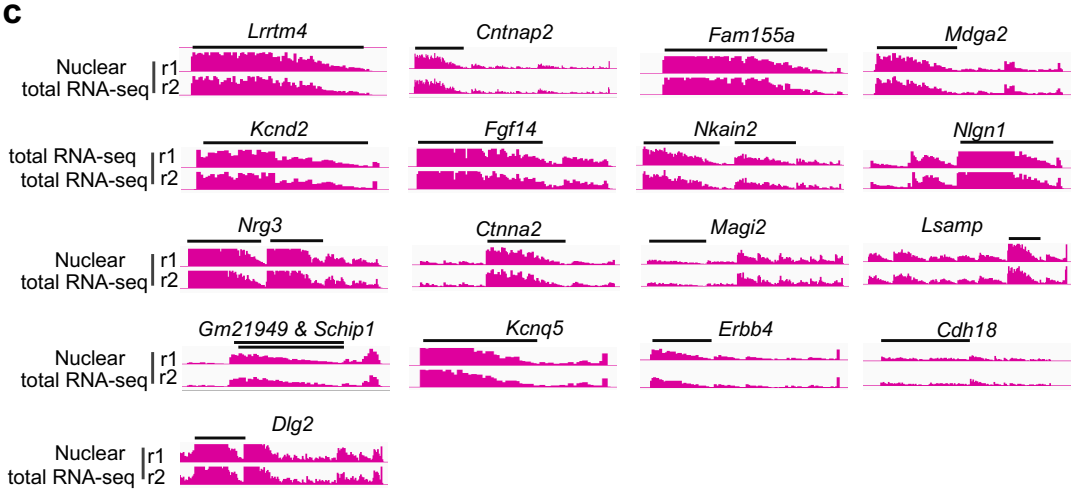
a

	Motif	Sequence of 20 nt upstream of AGGT	(T+C)%	T%	C%	G%	A%	3'SS MaxEnt score
chr16_+_40655810_Lsamp	AGGTAAGT	TTTCTTCTTTTCTCTCTTCC	1	0.65	0.35	0	0	14.81
chr9_+_47633924_Cadm1	AGGTAAGT	TTTTCTCTCCTCTTTCTTTT	1	0.7	0.3	0	0	13
chr13_+_109333510_Pde4d	AGGTAAGT	CTCATCTCTCTTTCTCTTTT	0.95	0.6	0.35	0	0.05	14.12
chr13_+_109614417_Pde4d	AGGTAAGT	TGCTTTTCTTTTTCTTTTTC	0.95	0.75	0.2	0.05	0	15.18
chr16_+_41206970_Lsamp	AGGTAAGT	TTTTTTTTCATTCTCTCCTT	0.95	0.7	0.25	0	0.05	12.87
chr16_+_40262266_Lsamp	AGGTAAGT	TTTTCTATTTTTTTTCTTCT	0.95	0.8	0.15	0	0.05	12.03
chr9_+_28029505_Opcml	AGGTAAGT	CCCTTTCTTTGTCTTTCCCT	0.95	0.55	0.4	0.05	0	13.71
chr10_+_69595048_Ank3	AGGTAAGT	TTTCTCTCTTTTTCTTTTAC	0.95	0.7	0.25	0	0.05	14.57
chr9_-_29674403_Ntm	AGGTAAGT	TCTCCCGTCTCTCTTTTTAT	0.9	0.55	0.35	0.05	0.05	12.55
chr16_+_40979498_Lsamp	AGGTAAGT	TGTTGTTTCTTTTTTCTTTC	0.9	0.75	0.15	0.1	0	13.26
chr16_-_74151682_Robo2	AGGTAAGT	TGGCTCTTCATTTCTTCTTC	0.85	0.55	0.3	0.1	0.05	10.73
chr2_+_179627792_Cdh4	AGGTAAGT	AAACCTTTCCTCTTATTCCT	0.8	0.45	0.35	0	0.2	10.26
chr11_-_33718264_Kcnip1	AGGTAAGT	ACTTCTGTGTTCTTTCTTGC	0.8	0.55	0.25	0.15	0.05	13.77
chr13_+_109550007_Pde4d	AGGTAAGT	TTTGTGTGTTTTGTTTTTTT	0.8	0.8	0	0.2	0	12.59
chr14_+_119537648_Hs6st3	AGGTAAGT	TGACTTCTGTCCCATATCTC	0.75	0.4	0.35	0.1	0.15	9.3
chr16_-_67142933_Cadm2	AGGTAAGC	TTTTGTTTCCTTTTATTTTT	0.9	0.8	0.1	0.05	0.05	11.73
chr16_-_67364247_Cadm2	AGGTGAGT	CTCCCCCTTCTTGTTTTTAT	0.9	0.55	0.35	0.05	0.05	12.33
chr9_-_49642257_Ncam1	AGGTAAGG	TATCTCACCTCACAACCAAA	0.6	0.2	0.4	0	0.4	3.15
chr9_-_49719364_Ncam1	AGGTAAGG	TCACTCTGCCCTGTTAAAAC	0.65	0.3	0.35	0.1	0.25	7.31
chrX_-_51473338_Hs6st2	AGGTAAGG	CTCTGGCCATCTGGACACTC	0.65	0.25	0.4	0.2	0.15	2.98

b

	Motif	(T+C)%	T%	C%	G%	A%	3'SS MaxEnt score
chr10:69590385 (Ank3)	AGGTAAGT	0.9	0.7	0.2	0.05	0.05	11.64
Chr7:91187317 (Dlg2)	AGGTGAGT	0.8	0.45	0.35	0.15	0.05	12.81
Chr10:32590379 (Nkain2)	AGGTGAGT	0.75	0.55	0.2	0.15	0.1	10.50
Chr6:21365801 (Kcnd2)	AGGTGAGT	0.75	0.6	0.15	0.05	0.2	9.44

c



d

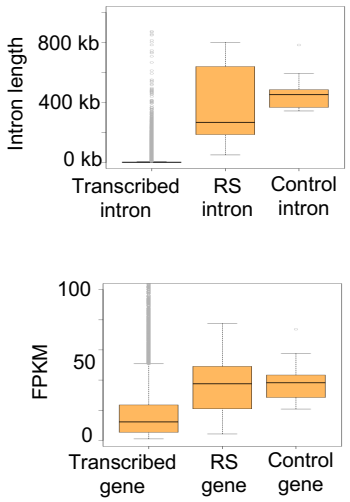
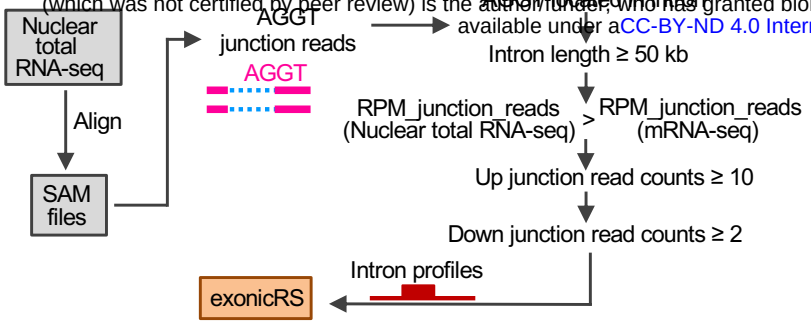
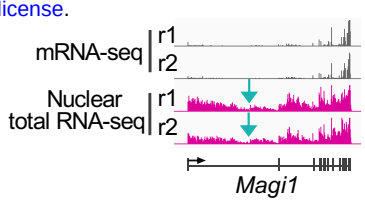


Figure S2: Primary sequence context of RS sites. (**a**) The sequence motifs, nucleotide percentages, and 3'SS MaxEnt scores of the 20 RS sites. The top 17 RS sites passed our prediction criteria of RS sites, while the bottom three RS sites failed. (**b**) The information of the non-RS AGGTs that passed our prediction criteria of RS sites. The top non-RS AGGT site resides in the RS intron of *Ank3*, while the bottom three non-RS AGGT sites reside in the three control long introns. (**c**) Profiles of nuclear total RNA-seq at the host genes of the 20 control long introns. Black bars indicate the control long introns. (**d**) Boxplots of the intron lengths (up) and host-gene expression levels (down) of introns transcribed in the mouse cerebral cortex (transcribed intron), RS introns, and control long introns.

a



c



b

Loci	Strand	Gene	Upstream exon ending	Number of Junction reads		r5'SS		Down 5'SS		Classification
				Up	Down	Sequence	MaxEnt score	Sequence	MaxEnt score	
chr2:17526050-17526130	-	Nebl	chr2:17529910	51	154	AATGTGGAA	-11.32	ATGGTAAGT	11.01	exonicRS
chr2:97522122-97522199	+	Lrrc4c	chr2:97468420	59	68	ACAGTGGAT	-7.51	AAGGTTGGA	5.64	exonicRS
chr7:61241310-61241445	-	A230006K03Rik	chr7:61282776	91	51	AAGGTGTTT	2.06	CAGGTTTGT	7.44	exonicRS
chr12:46766290-46766391	-	Nova1	chr12:46816886	61	50	CAGGTGCTG	4.87	ACTGTAAGT	8.59	exonicRS
chr5:57785821-57785880	+	Pcdh7	chr5:57722278	62	41	CAGGTGGAT	3.56	AATGTATGT	5.87	exonicRS
chr1:123595259-123595324	-	Dpp10	chr1:123650288	25	31	CAGGTGCTT	4.83	GAGGTAACA	8.1	exonicRS
chr1:124638974-124639139	-	Dpp10	chr1:124845719	43	34	AAGGTGTCT	2.57	CAGGTACGG	10.88	exonicRS
chr12:44527672-44527848	+	Nrcam	chr12:44456702	55	31	AAGGTTCCT	0.89	TTGGTATAT	1.75	exonicRS
chr3:158361404-158361562	-	Lrrc7	chr3:158413130	27	31	CAAGTCTCC	-11.96	AAGGTGAGA	8.68	exonicRS
chr7:75224697-75224831	-	Sv2b	chr7:75308205	43	25	AAGGTCTGG	3.93	CTGGTCAGT	6.6	exonicRS
chr2:136200589-136200635	-	Pak7	chr2:136269323	23	23	AAAGTTTCT	-9	AAGGTAAAA	8.38	exonicRS
chrX:77584794-77584859	+	Tbl1x	chrX:77511269	27	20	GAGGTAATT	8.88	TGAGTAAGT	8.82	exonicRS
chr3:159150266-159150351	+	Gm20752	chr3:159090393	21	19	AAGGTACAG	8.04	TTAGTAAGT	7.79	exonicRS
chr2:7282441-7282578	-	Celf2	chr2:7395858	24	18	CTGGTCTTT	-3.78	AAAGTAAGT	9.72	exonicRS
chr7:64506937-64507022	+	Apba2	chr7:64502253	17	17	CAGGTCTCC	0.85	GAGGTAGGA	8.24	exonicRS
chr16:6158060-6158206	+	Rbfox1	chr16:5885864	12	14	TAGGTGCAG	0.29	AAAGTAAGT	9.72	exonicRS
chr6:94050641-94050766	-	Magi1	chr6:94283010	41	13	AAGGTAAGG	10.51	GTTGTAAGT	8.3	RS & exonicRS
chr3:29170414-29170593	+	Egfem1	chr3:29153601	18	13	ATGGTTAAA	-2.37	AAGGTAAAG	9.06	exonicRS
chr8:54881710-54881841	-	Wdr17	chr8:54886907	38	12	AAGGTGCTT	4.72	CAGGTACCA	7.88	exonicRS
chr8:50161927-50162063	-	Gm45341	chr8:50415134	10	11	AACGTGTTC	-7.75	CTGGTGAGT	10.1	exonicRS
chr7:96232099-96232201	+	Tenm4	chr7:96211990	15	9	ACTGTGAAG	-5.79	ACTGTAAGG	5.19	exonicRS
chr7:84005117-84005995	-	Cemip	chr7:84086258	23	2	GAGGTGACA	3.85	CAGGTCCGA	4.4	exonicRS

Loci	Strand	Gene	Upstream exon ending	Number of Junction reads		r5'SS		Down 5'SS		Classification
				Up	Down	Sequence	MaxEnt score	Sequence	MaxEnt score	
chr16:67364176-67364248	-	Cadm2	chr16:67620171	317	22	AAGGTGAGT	10.47	TGGGTAAGT	10.24	RS & exonicRS
chr16:67142893-67142934	-	Cadm2	chr16:67620171	219	22	AAGGTAAGC	10.22	AAGGTAAAA	8.38	RS & exonicRS
chr16:41206972-41207028	+	Lsamp	chr16:40786060	100	5	CAGGTAAGT	10.86	GAGGCAAGT	3.32	RS & exonicRS
chr9:29674281-29674404	-	Ntm	chr9:29962845	475	5	AAGGTAAGT	11	CAGGTAGGT	10.28	RS & exonicRS
chr16:40262268-40262469	+	Lsamp	chr16:39984727	161	5	AAGGTAAGT	11	AAGGTAAGT	11	RS & exonicRS
chr10:69595050-69595128	+	Ank3	chr10:69534284	173	4	AAGGTAAGT	11	AAGGTTCGT	8.44	RS & exonicRS
chr16:40655812-40655859	+	Lsamp	chr16:39984727	79	4	AAGGTAAGT	10.86	GTGGTAAGT	10.36	RS & exonicRS
chr9:47633926-47634076	+	Cadm1	chr9:47530505	62	4	CAGGTAAGT	10.86	CAGGTAATT	8.55	RS & exonicRS
chr16:40979500-40979631	+	Lsamp	chr16:40786060	112	3	CAGGTAAGT	11	CAGGTGCTT	4.83	RS & exonicRS
chr9:28029507-28029642	+	Opcml	chr9:27791446	334	1	CAGGTAAGT	10.86	AAGGTATGG	9.26	RS & exonicRS
chr14:119537650-119537790	+	Hs6st3	chr14:119139118	202	0	CAGGTAAGT	10.86	TTGGTGACA	-1.41	RS
chr13:109550009-109550238	+	Pde4d	chr13:109442353	42	0	ACGGTAAGT	11.81	CAAGTACAA	-1.09	RS
chr13:109614419-109614568	+	Pde4d	chr13:109442353	33	0	ACGGTAAGT	11.81	TTGGTTTGG	-1.65	RS
chr2:179627794-179627953	+	Cdh4	chr2:179444811	25	0	AAGGTAAGT	11	TTGGTCTGT	0.67	RS
chr9:49642143-49642258	-	Ncam1	chr9:49798678	135	0	CAGGTAAGG	11.08	AAAGTGTGA	-1.44	RS
chr13:109333512-109333719	+	Pde4d	chr13:109117080	128	0	ACGGTAAGT	11.81	AAAGTTGGA	-3.21	RS
chr16:74151336-74151683	-	Robo2	chr16:74352551	97	0	CATGTAAGT	8.31	TTGGTTATA	-5.68	RS
chr9:49719316-49719365	-	Ncam1	chr9:49798678	83	0	CAGGTAAGG	11.08	AAGGTTTCT	0.28	RS
chrX:51473121-51473339	-	Hs6st2	chrX:51679890	26	0	CAGGTAAGG	11.08	ATGGTTTTA	-5.54	RS
chr11:33718032-33718265	-	Kcnip1	chr11:33843130	8	0	AAGGTAAGT	11	CTGGTGGCC	-2.74	RS

Figure S3: Identification of exonicRS. (**a**) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify exonicRS. (**b**) Tables of information of genomic loci, junction reads, 5'SS sequences, 5'SS MaxEnt scores, and classifications of exonicRS (up) and RS sites (down). (**c**) Sequencing profile at *Magi1* locus. Green arrows indicate the RS AGGT loci.