# 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

70 context distinguishes RS AGGT sites from non-RS AGGT sites and determines the choice
71 between RS and exon-like RS. Our findings provide novel insights into the mechanism
72 and evolution of RS in long genes.
73
74 **Results**
75 **Nuclear total RNA is enriched for nascent transcripts and RS splicing**
76 **intermediates.** Two features are critical for the identification of RS sites from total RNA-
77 seq data (Fig. 1a). One feature is the saw-tooth pattern of decreasing read density in the
78 host intron (RS intron)[3, 4] (Fig. 1a, red triangles), which indicates nascent transcripts
79 and splicing status[13, 14]. The other feature is the junction reads spanning upstream
80 exon and the RS AGGT site (RS junction reads, Fig. 1a), which indicate the RS splicing
81 intermediates. Given that nascent transcripts and RS splicing intermediates are localized
82 in the nucleus, we asked whether nuclear total RNA is enriched for nascent transcripts
83 and RS splicing intermediates compared to whole-cell total RNA (Fig. 1b). To answer this
84 question, we reanalyzed the whole-cell total RNA-seq and nuclear total RNA-seq data
85 (Figure S1a in Additional file 1) that we generated from the cerebral cortex of 6-week-old
86 mice[11, 12].
87 We first examined nascent transcripts in whole-cell total RNA-seq data, nuclear total
88 RNA-seq data, and poly(A) enriched messenger RNA-seq (mRNA-seq) data from the
89 mouse cerebral cortex[15]. We calculated the proportions of reads mapped to introns
90 because total RNA-seq reads in introns indicate nascent transcripts[13, 14]. We found
91 that 77% of the uniquely mapped reads from the nuclear total RNA-seq data were
92 localized in introns (Figure S1a,b in Additional file 1), which was significantly higher than

93 the 41% from the whole-cell total RNA-seq data ($P < 2.2^{-16}$, one-tailed Fisher's Exact
94 Test). In contrast, only 23% of the uniquely mapped reads were localized in introns from
95 mRNA-seq data (Figure S1a in Additional file 1). Thus, nuclear total RNA-seq captures
96 more nascent transcripts than whole-cell total RNA-seq does.

97 We next compared the two RS features (Fig. 1a) between the two total RNA-seq
98 methods in a known RS intron, the intron of *Hs6st3[4]*. A more distinct saw-tooth pattern
99 was observed in nuclear total RNA-seq data than that in whole-cell total RNA-seq data
100 (Fig. 1c). Furthermore, the number of RS junction reads at *Hs6st3* was 3-fold higher in
101 nuclear total RNA-seq data than in whole-cell total RNA-seq data (Fig. 1d). Together,
102 these results demonstrate that nuclear total RNA is enriched for nascent transcripts and
103 RS splicing intermediates.

104

105 **Identification of RS sites from nuclear total RNA-seq data.** To identify new RS sites,
106 we developed a pipeline based on the RS junction reads and the saw-tooth patterns in
107 RS introns from total RNA-seq data (Fig. 1e and Figure S1c in Additional file 1). We
108 extracted all junction reads spanning AGGT sites from the alignment results, focused on
109 AGGT sites located in long introns (length ≥ 50 kb), selected sites enriched for RS junction
110 reads compared to mRNA-seq, and chose sites containing 10 or more RS junction reads
111 as RS site candidates (Figure S1c in Additional file 1). The candidates were further refined
112 based on saw-tooth patterns in their host introns (Figure S1c in Additional file 1).

113 By applying this pipeline to our nuclear total RNA-seq data, we identified 19 RS sites,
114 which include all of the ten known RS sites in mice[4] (Figure S1d in Additional file 1),
115 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

139 also identified seven and five junction reads for them in the female data (Figure S1d in
140 Additional file 1), although they failed to pass our criteria of 10 junction reads. Together,
141 these results indicate that RS is constitutive between male and female excitatory neurons
142 in the mouse cerebral cortex.
143
144 **Characteristics of the RS sites in mice.** We next investigated the characteristics of the
145 20 RS sites we identified in mice (Figure S1d in Additional file 1). Given the high
146 conservation of RS mechanism among species[3, 4, 7-9], we first investigated the
147 sequence conservation of the 600 nt regions surrounding the RS sites. In agreement with
148 previous studies[3, 4, 7, 8], the AGGT motif at the RS sites is highly conserved across 60
149 vertebrate genomes, showing high phylogenetic p-value scores (phyloP scores)
150 compared to upstream and downstream regions (Fig. 3a). The RS introns, showing a
151 median length of 267 kb and a mean length of 382 kb, were significantly longer than
152 introns transcribed in the mouse cerebral cortex (Fig. 3b). Based on the profile of histone
153 H3 lysine 4 trimethylation in the mouse cerebral cortex[12], we found that 75% of RS
154 introns were the first intron, and the rest 25% were the second intron in their host genes
155 (Fig. 3c).
156
157 **RS genes are specifically expressed in the brain.** RS genes are highly expressed long
158 genes, showing a median length of 628 kb (Figure S1e in Additional file 1). To determine
159 whether RS genes are specifically expressed in the brain, we investigated the expression
160 patterns of RS genes among 22 mouse tissues from the ENCODE project[15]. We found
161 that RS genes are specifically expressed in brain regions including the cortex, frontal
162 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

186 composition of these regions and found the high percentages of thymine and cytosine in
187 RS AGGT sites, but not in non-RS AGGT sites (Fig. 4d). Lastly, given that RS AGGT sites
188 are used as the 3' splice sites (3'SS) during splicing, we quantified the strengths of 3'SS
189 using MaxEntScan[20] and found significantly higher 3'SS scores of RS AGGT sites than
190 that of non-RS AGGT sites (Fig. 4e). Together, these results demonstrate that RS AGGT
191 sites exhibit specific primary sequence features compared to non-RS AGGT sites.
192
193 **Computational prediction of RS sites based on primary sequences alone.** Next, we
194 investigated whether primary sequence features alone could be used to predict RS sites.
195 We developed a computational pipeline to predict RS sites based on the presence of
196 AGGT motifs and three additional sequence features (Fig. 4f, blue box). First, the AGGT
197 site should have a U1 snRNA binding site, AGGTAAGB (B is T, G, or C) or AGGTGAGT
198 (Fig. 4f). Second, the 20 nt region upstream of the AGGT site should have a
199 polypyrimidine tract. The percentage of thymine in this region should be ≥ 40%, and the
200 combined percentage of thymine and cytosine in this region should be ≥ 75% (Fig. 4f).
201 Third, the AGGT site should have a strong 3'SS. The MaxEnt score of 3'SS should be ≥
202 9.3 (Fig. 4f). The cutoff values of these criteria were obtained from the values of *Hs6st3*
203 (Figure S2a in Additional file 2).
204 We applied the computational prediction pipeline on the 23,710 AGGT sites in the 20
205 RS introns. Eighteen AGGTs sites passed the criteria (Fig. 4f), which include 85% (17 of
206 20) of RS AGGT sites that we identified from the sequencing data (Fig. 4f and Figure S2a
207 in Additional file 2). In contrast, 99.996% (all but one of the 23,690) of non-RS AGGT sites
208 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.

232 5a). By applying this pipeline to our nuclear total RNA-seq data, we discovered 22
233 exonicRS in the introns of 21 long genes (Figure S3b in Additional file 3).
234 To illuminate the features of exonicRS, we first analyzed the numbers of junction
235 reads in RS and exonicRS. We found that RS exhibited 37-fold more Up junction reads
236 than Down junction reads (Fig. 5b). In contrast, exonicRS exhibited comparable numbers
237 of Up and Down junction reads (Fig. 5b). Next, we investigated the 5' splice sites (5'SS)
238 at the two ends of the exons at RS sites and exonicRS, the reconstituted 5'SS (r5'SS)
239 after the first step of splicing and the downstream 5'SS (Down 5'SS). These were
240 investigated because the competition between them may be associated with the inclusion
241 of the RS exons in mature transcripts[4, 7, 9]. We examined the sequence features of
242 r5'SS and Down 5'SS using WebLogo[17] and found that RS showed an enrichment of
243 5'SS motif (AGGTAAGT) at the r5'SS but not at the Down 5'SS, while exonicRS showed
244 the opposite trend (Fig. 5c). Quantification of the strengths of r5'SS and Down 5'SS using
245 MaxEntScan[20] demonstrated that the MaxEnt scores of r5'SS were significantly higher
246 than that of Down 5'SS in RS sites (Fig. 5d), while the MaxEnt scores of r5'SS were
247 significantly lower than that of Down 5'SS in exonicRS (Fig. 5d). Taken together, these
248 results demonstrate that RS and exonicRS exhibit distinct molecular features.
249
250 **Strengths of 5'SS are able to distinguish RS and exonicRS.** We next asked whether
251 the strengths of r5'SS and Down 5'SS were able to distinguish RS and exonicRS. We
252 plotted the MaxEnt scores of 5'SS of the 20 RS and the 22 exonicRS (Fig. 5e).
253 Unexpectedly, they were classified into three categories based on the 5'SS scores (Figure
254 5e and Figure S3b in Additional file 3). One category is exonicRS, which exhibits high

255 Down 5'SS scores and low r5'SS scores (Fig. 5e, green triangles). Another category is
256 RS, which exhibits high r5'SS scores and low Down 5'SS scores (Fig. 5e, pink dots).
257 There is also a third category that contains ten RS and one exonicRS (Fig. 5e, blue oval
258 region). Further analyses revealed that these ten RS sites in the third category all
259 contained Down junction reads (Figure S3b in Additional file 3), suggesting that the third
260 category is a combination of RS and exonicRS (Fig. 5a). This combination is further
261 supported by the observation that the exonicRS in the third category exhibits a weak saw-
262 tooth pattern, which failed to pass our stringent criteria in our initial identification of RS
263 sites (Figure S3c in Additional file 3). Notably, both of the two RS sites (*Cadm1* and *Ank3*),
264 which have been experimentally confirmed to utilize the exon definition mechanism in
265 mammals[4], were classified into the third category (Figure S3b in Additional file 3).
266 Together, these results support that the strengths of 5'SS are able to distinguish RS and
267 exonicRS (Fig. 5f).
268

269 **Discussion**

270 A long gene consumes more time and resources to make a transcript than a short
271 gene does. To overcome the length constraint, long genes may use specific mechanisms
272 to regulate their transcription and co-transcriptional processes, such as splicing. RS is a
273 splicing mechanism specific to long introns in long genes. In this study, we developed an
274 efficient pipeline to identify RS from nuclear total RNA-seq data, investigated the primary
275 sequence context of RS, and discovered a novel type of RS events. Our identification of
276 20 RS sites from high-depth nuclear total RNA-seq data suggests that RS is likely a
277 special splicing mechanism only for a small portion of introns in the mammalian genome.

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

281 Each long intron contains hundreds to thousands of AGGT sites, suggesting that
282 AGGT motif alone could not determine the RS mechanism. Previous studies reported that
283 RS AGGT sites exhibit specific sequence context, but it remains unclear whether the
284 sequence context alone could predict RS. We systematically examined the sequence
285 context of RS sites and developed a computational pipeline to predict RS with high
286 accuracy and low false discovery rate. Our findings indicate that RS is largely determined
287 by the primary sequence context. Notably, several RS sites failed our prediction pipeline,
288 suggesting that other factors, such as RNA binding proteins and RNA structures[3, 29-
289 34], may also play a role in the definition of RS mechanism. Thus, further investigations
290 are necessary to integrate the primary sequence context, RNA binding protein binding,
291 and RNA structures to illustrate the molecular basis of RS.

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

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

302 RS genes are specifically expressed in the brain and are genetically linked to various
303 human brain disorders. For example, *ANK3* encodes ankyrin-G and is linked to autism
304 spectrum disorders, attention deficit hyperactivity disorder, intellectual disability, and
305 bipolar disorder[36-39]. Also, *NTM* encodes neurotrimin and is linked to autism spectrum
306 disorders and attention deficit hyperactivity disorder[40, 41]. The *PDE4D* encodes
307 phosphodiesterase 4D and is linked to schizophrenia, psychosis, acrodysostosis, and
308 neuroticism[42-44]. Notably, PDE4D Inhibitors are in clinical trials for the treatment of
309 Alzheimer's disease and Fragile X syndrome[45, 46]. Given that the disruption of the RS
310 process interfered with the RS gene function and caused abnormality in the central
311 nervous system[7], further investigation will be necessary to illuminate whether RS
312 mechanism contributes to the pathophysiology of these human brain disorders.

313

314 **Conclusions**

315 In this study, we reveal the molecular mechanism of RS in long introns of long genes. Our
316 results highlighted that nuclear total RNA-seq is an efficient approach to investigate RS.
317 We develop a novel pipeline to identify RS events, characterize the cell type specificity
318 and genomic features of RS sites, and discover a novel type of RS events. Through
319 analysis of primary sequence, we demonstrate that RS is largely determined by the
320 primary sequence context, thus providing novel insights into the RS mechanism. Our
321 discovery of exonicRS indicates a new mechanism by which long genes could acquire
322 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-

346 read-specific sam files. These sam files were further converted into bam files using
347 samtools. The junction-read-specific bam files were loaded into IGV for visualization.
348 **Junction reads spanning AGGT sites.** All AGGT sites (20,403,114) in the mouse mm10
349 genome were identified, and only AGGT sites (4,767,575) located in the gene sense
350 regions were kept for further analysis. The AGGT sites that were kept were used to screen
351 the junction-read-specific sam files. The numbers of junction reads spanning each AGGT
352 site (joining the upstream exon and sequences following GT) were counted. The counts
353 of junction reads were further normalized to the sequencing depth to obtain the RPM
354 values.
355 **Pipeline to identify RS sites.** The schematic of this pipeline is shown in Figure S1c in
356 Additional file 1. Briefly, AGGT sites located in introns longer than or equal to 50 kb were
357 extracted. Sites that showed a larger RPM value of junction reads in total RNA-seq data
358 than in mRNA-seq data were kept for downstream analyses. The counts of junction reads
359 of biological replicates were merged, and AGGT sites that contained 10 or more junction
360 reads were identified as RS site candidates. The RS site candidates were further refined
361 as RS sites if the host intron showed a clear saw-tooth pattern.
362 **FPKM of nuclear total RNA-seq data.** The number of reads mapped to the exonic
363 regions of each gene were calculated to get the raw counts. The raw counts were then
364 normalized to the exon lengths of that gene and to the sequencing depth of that data set
365 to get the FPKM values.
366 **Phylogenetic p-value (phyloP) scores.** The phyloP scores, which were calculated by
367 the PHAST package for multiple alignments of 59 vertebrate genomes to the mouse

368 genome, were obtained from the UCSC Genome Browser
369 (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phyloP60way/).
370 **Gene expression profiles in 22 mouse tissues.** The expression profiles (FPKM values)
371 of RS genes in 22 mouse tissues were obtained from the LongGeneDB database
372 (https://longgenedb.com).
373 **Control long introns.** Introns transcribed in the mouse cerebral cortex were sorted by
374 intron lengths. Only the highly expressed introns (host gene FPKM > 20) were retained,
375 and the top 20 longest introns were used as the control long introns.
376 **Single nucleus RNA-seq data.** The expression levels of RS genes in the 24 cell types
377 in the mouse cerebral cortex were obtained from the LongGeneDB database
378 (https://longgenedb.com).
379 **WebLogo analysis.** WebLogo 3[17] (http://weblogo.threeplusone.com) was used to
380 perform the sequence logo analysis. The Output Format was chosen as “PNG (high res.)”,
381 and the Stacks per Line was set to “80”. The default values were used for other
382 parameters.
383 **MaxEntScan 3’SS analysis.** The 3’SS scores were calculated by MaxEntScan::score3ss
384 (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html)[20]. The
385 input sequences were composed of the 18 nt region upstream of the AGGT, the AGGT
386 motif, and the one nucleotide following AGGT (18nt + AGGT + 1nt). The three models -
387 Maximum Entropy Model, First-order Markov Model, and Weight Matrix Model – were
388 selected. The MaxEnt scores were used as the 3’SS scores.

389 **Reconstituted 5' splice sites (r5'SS).** The r5'SS sequences were composed of the last
390 30 nucleotides of the upstream exon, the GT motif, and the 20 nucleotides following
391 AGGT (30nt + GT + 20nt).
392 **MaxEntScan 5'SS analysis.** The 5'SS scores were calculated by MaxEntScan::score5ss
393 (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)[20]. The input
394 sequences for 5'SS were composed of three nucleotides before the GT, the GT motif,
395 and the four nucleotides following AGGT (3nt + GT + 4nt). The input sequences of r5'SS
396 and Down 5'SS were listed in Figure S3b in Additional file 3.
397 **Pipeline to identify exonicRS.** The schematic of this pipeline is shown in Figure S3a in
398 Additional file 3. Briefly, AGGT sites located in introns longer than or equal to 50 kb were
399 extracted. The AGGT sites that showed a larger RPM value of junction reads in total RNA-
400 seq data than in mRNA-seq data were kept for downstream analyses. The counts of the
401 junction reads of the biological replicates were merged. AGGT sites contained 10 or more
402 Up-junction reads and two or more Down-junction reads were identified as exonicRS
403 candidates. The exonicRS candidates were further refined as exonicRS if the host intron
404 showed an exon-like but not saw-tooth like pattern.
405
406 **Abbreviations:**
407 RS: recursive splicing; mRNA: messenger RNA; kb: kilobase; RNA-seq: RNA
408 sequencing; mRNA-seq: poly(A) enriched messenger RNA sequencing; phyloP:
409 phylogenetic p-value; 3'SS: 3' splice site; 5'SS; 5' splice site; RPM, reads per million
410 uniquely mapped reads.
411

412 **Declarations:**

413 **Ethics approval and consent to participate**

414 Not applicable.

415 **Consent for publication**

416 Not applicable.

417 **Availability of data and materials**

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

422 **Competing interests**

423 The authors declare that they have no competing interests.

424 **Funding**

425 None.

426 **Authors' contributions**

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

430

431 **Acknowledgements**

432 We thank Dr. Raddy Ramos, Dr. Weikang Cai, Dr. Lars Udo-Bellner, and members of the
433 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.

469 12. Zhao YT, Kwon DY, Johnson BS, Fasolino M, Lamonica JM, Kim YJ, Zhao BS, He C, Vahedi
470 G, Kim TH *et al*: **Long genes linked to autism spectrum disorders harbor broad enhancer-**
471 **like chromatin domains**. *Genome Res* 2018, **28**(7):933-942.
472 13. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L: **Total**
473 **RNA sequencing reveals nascent transcription and widespread co-transcriptional**
474 **splicing in the human brain**. *Nat Struct Mol Biol* 2011, **18**(12):1435-1440.
475 14. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder
476 M, Gingeras TR, Guigo R: **Deep sequencing of subcellular RNA fractions shows splicing**
477 **to be predominantly co-transcriptional in the human genome but inefficient for**
478 **lncRNAs**. *Genome Res* 2012, **22**(9):1616-1625.
479 15. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J,
480 Zaleski C, See LH *et al*: **Enhanced transcriptome maps from multiple mouse tissues reveal**
481 **evolutionary constraint in gene expression**. *Nat Commun* 2015, **6**:5903.
482 16. Hu P, Fabyanic E, Kwon DY, Tang S, Zhou Z, Wu H: **Dissecting Cell-Type Composition and**
483 **Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel**
484 **Single-Nucleus RNA-Seq**. *Mol Cell* 2017, **68**(5):1006-1015 e1007.
485 17. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**.
486 *Genome Res* 2004, **14**(6):1188-1190.
487 18. Seraphin B, Kretzner L, Rosbash M: **A U1 snRNA:pre-mRNA base pairing interaction is**
488 **required early in yeast spliceosome assembly but does not uniquely define the 5'**
489 **cleavage site**. *EMBO J* 1988, **7**(8):2533-2538.
490 19. Fu XY, Ge H, Manley JL: **The role of the polypyrimidine stretch at the SV40 early pre-**
491 **mRNA 3' splice site in alternative splicing**. *EMBO J* 1988, **7**(3):809-817.
492 20. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with**
493 **applications to RNA splicing signals**. *J Comput Biol* 2004, **11**(2-3):377-394.
494 21. Robberson BL, Cote GJ, Berget SM: **Exon definition may facilitate splice site selection in**
495 **RNAs with multiple exons**. *Mol Cell Biol* 1990, **10**(1):84-94.
496 22. Georgomanolis T, Sofiadis K, Papantonis A: **Cutting a Long Intron Short: Recursive**
497 **Splicing and Its Implications**. *Front Physiol* 2016, **7**:598.
498 23. Hollander D, Naftelberg S, Lev-Maor G, Kornblihtt AR, Ast G: **How Are Short Exons**
499 **Flanked by Long Introns Defined and Committed to Splicing?** *Trends Genet* 2016,
500 **32**(10):596-606.
501 24. Pai AA, Henriques T, McCue K, Burkholder A, Adelman K, Burge CB: **The kinetics of pre-**
502 **mRNA splicing in the Drosophila genome and the influence of gene architecture**. *Elife*
503 2017, **6**.
504 25. Shepard S, McCreary M, Fedorov A: **The peculiarities of large intron splicing in animals**.
505 *PLoS One* 2009, **4**(11):e7853.
506 26. Singh J, Padgett RA: **Rates of in situ transcription and splicing in large human genes**. *Nat*
507 *Struct Mol Biol* 2009, **16**(11):1128-1133.
508 27. Swinburne IA, Miguez DG, Landgraf D, Silver PA: **Intron length increases oscillatory**
509 **periods of gene expression in animal cells**. *Genes Dev* 2008, **22**(17):2342-2346.
510 28. Fong YW, Zhou Q: **Stimulatory effect of splicing factors on transcriptional elongation**.
511 *Nature* 2001, **414**(6866):929-933.

512 29. Blazquez L, Emmett W, Faraway R, Pineda JMB, Bajew S, Gohr A, Haberman N, Sibley CR,
513 Bradley RK, Irimia M *et al*: **Exon Junction Complex Shapes the Transcriptome by**
514 **Repressing Recursive Splicing**. *Mol Cell* 2018, **72**(3):496-509 e499.
515 30. Patton RD, Sanjeev M, Woodward LA, Mabin JW, Bundschuh R, Singh G: **Chemical**
516 **crosslinking enhances RNA immunoprecipitation for efficient identification of binding**
517 **sites of proteins that photo-crosslink poorly with RNA**. *RNA* 2020.
518 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**
519 **situ profiling of RNA-RNA spatial interactions**. *Nature* 2020.
520 32. Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H *et al*: **Genome-**
521 **wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing**
522 **repressor to modulate exon inclusion or skipping**. *Mol Cell* 2009, **36**(6):996-1006.
523 33. Ule J, Blencowe BJ: **Alternative Splicing Regulatory Networks: Functions, Mechanisms,**
524 **and Evolution**. *Mol Cell* 2019, **76**(2):329-345.
525 34. Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, Park SS, Garcia KE,
526 Gelboin-Burkhart C, Nguyen TB *et al*: **Principles of RNA processing from analysis of**
527 **enhanced CLIP maps for 150 RNA binding proteins**. *Genome Biol* 2020, **21**(1):90.
528 35. Roy M, Kim N, Xing Y, Lee C: **The effect of intron length on exon creation ratios during**
529 **the evolution of mammalian genomes**. *RNA* 2008, **14**(11):2261-2273.
530 36. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek
531 AG, DiLullo NM, Parikshak NN, Stein JL *et al*: **De novo mutations revealed by whole-**
532 **exome sequencing are strongly associated with autism**. *Nature* 2012, **485**(7397):237-
533 241.
534 37. Iqbal Z, Vandeweyer G, van der Voet M, Waryah AM, Zahoor MY, Besseling JA, Roca LT,
535 Vulto-van Silfhout AT, Nijhof B, Kramer JM *et al*: **Homozygous and heterozygous**
536 **disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric**
537 **disorders**. *Hum Mol Genet* 2013, **22**(10):1960-1970.
538 38. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G,
539 Perlis RH, Green EK *et al*: **Collaborative genome-wide association analysis supports a role**
540 **for ANK3 and CACNA1C in bipolar disorder**. *Nat Genet* 2008, **40**(9):1056-1058.
541 39. Schulze TG, Detera-Wadleigh SD, Akula N, Gupta A, Kassem L, Steele J, Pearl J, Strohmaier
542 J, Breuer R, Schwarz M *et al*: **Two variants in Ankyrin 3 (ANK3) are independent genetic**
543 **risk factors for bipolar disorder**. *Mol Psychiatry* 2009, **14**(5):487-491.
544 40. Maruani A, Huguet G, Beggiato A, ElMaleh M, Toro R, Leblond CS, Mathieu A, Amsellem
545 F, Lemiere N, Verloes A *et al*: **11q24.2-25 micro-rearrangements in autism spectrum**
546 **disorders: Relation to brain structures**. *Am J Med Genet A* 2015, **167A**(12):3019-3030.
547 41. Brevik EJ, van Donkelaar MM, Weber H, Sanchez-Mora C, Jacob C, Rivero O, Kittel-
548 Schneider S, Garcia-Martinez I, Aebi M, van Hulzen K *et al*: **Genome-wide analyses of**
549 **aggressiveness in attention-deficit hyperactivity disorder**. *Am J Med Genet B*
550 *Neuropsychiatr Genet* 2016, **171**(5):733-747.
551 42. Sinha V, Ukkola-Vuoti L, Ortega-Alonso A, Torniainen-Holm M, Therman S, Tuulio-
552 Henriksson A, Jylha P, Kaprio J, Hovatta I, Isometsa E *et al*: **Variants in regulatory elements**
553 **of PDE4D associate with major mental illness in the Finnish population**. *Mol Psychiatry*
554 2019.

555 43. Lee H, Graham JM, Jr., Rimoin DL, Lachman RS, Krejci P, Tompson SW, Nelson SF, Krakow
556 D, Cohn DH: **Exome sequencing identifies PDE4D mutations in acrodysostosis**. *Am J Hum*
557 *Genet* 2012, **90**(4):746-751.
558 44. Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, Hettema JM, An SS, Neale
559 MC, van den Oord EJ *et al*: **A whole genome association study of neuroticism using DNA**
560 **pooling**. *Mol Psychiatry* 2008, **13**(3):302-312.
561 45. Pan T, Xie S, Zhou Y, Hu J, Luo H, Li X, Huang L: **Dual functional cholinesterase and PDE4D**
562 **inhibitors for the treatment of Alzheimer's disease: Design, synthesis and evaluation of**
563 **tacrine-pyrazolo[3,4-b]pyridine hybrids**. *Bioorg Med Chem Lett* 2019, **29**(16):2150-2152.
564 46. Gurney ME, Nugent RA, Mo X, Sindac JA, Hagen TJ, Fox D, 3rd, O'Donnell JM, Zhang C, Xu
565 Y, Zhang HT *et al*: **Design and Synthesis of Selective Phosphodiesterase 4D (PDE4D)**
566 **Allosteric Inhibitors for the Treatment of Fragile X Syndrome and Other Brain Disorders**.
567 *J Med Chem* 2019, **62**(10):4884-4901.
568 47. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, Harrison
569 PW, Holt S, Hussein A, Ivanov E *et al*: **The European Nucleotide Archive in 2019**. *Nucleic*
570 *Acids Res* 2020, **48**(D1):D70-D76.
571 48. Dobin A, Gingeras TR: **Mapping RNA-seq Reads with STAR**. *Curr Protoc Bioinformatics*
572 2015, **51**:11 14 11-11 14 19.
573 49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
574 Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**.
575 *Bioinformatics* 2009, **25**(16):2078-2079.
576 50. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F,
577 Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis**.
578 *Nucleic Acids Res* 2016, **44**(W1):W160-165.
579 51. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP:
580 **Integrative genomics viewer**. *Nat Biotechnol* 2011, **29**(1):24-26.
581 52. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean
582 IM, Azov AG, Bennett R *et al*: **Ensembl 2020**. *Nucleic Acids Res* 2020, **48**(D1):D682-D688.
583

584

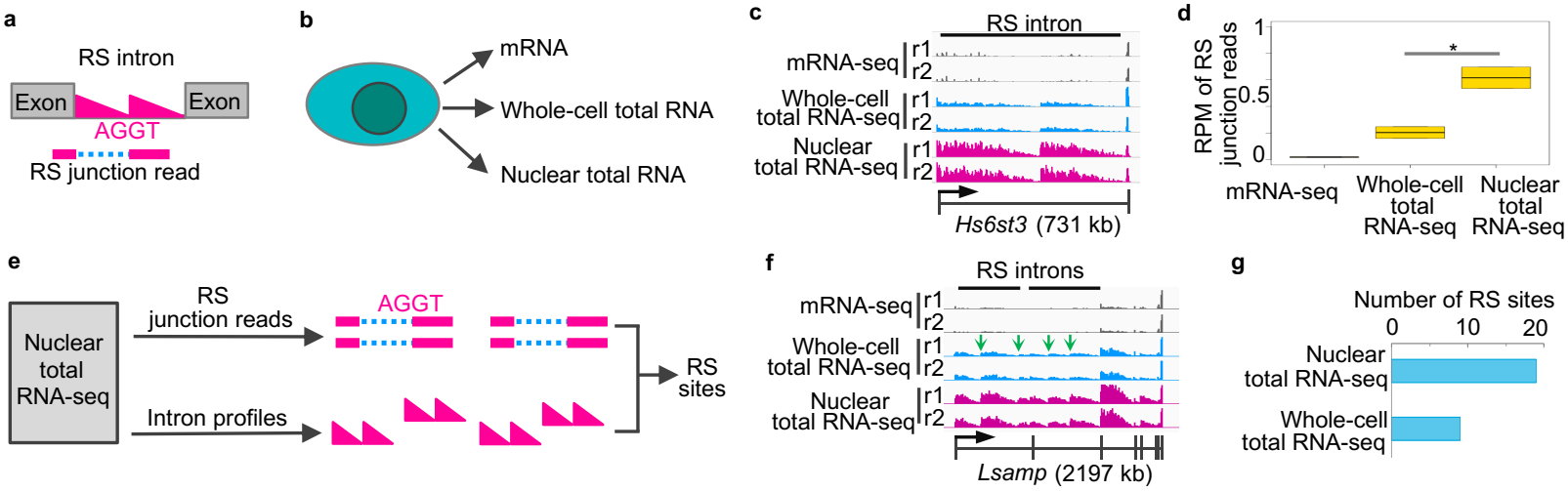
585 **Figure Legends**

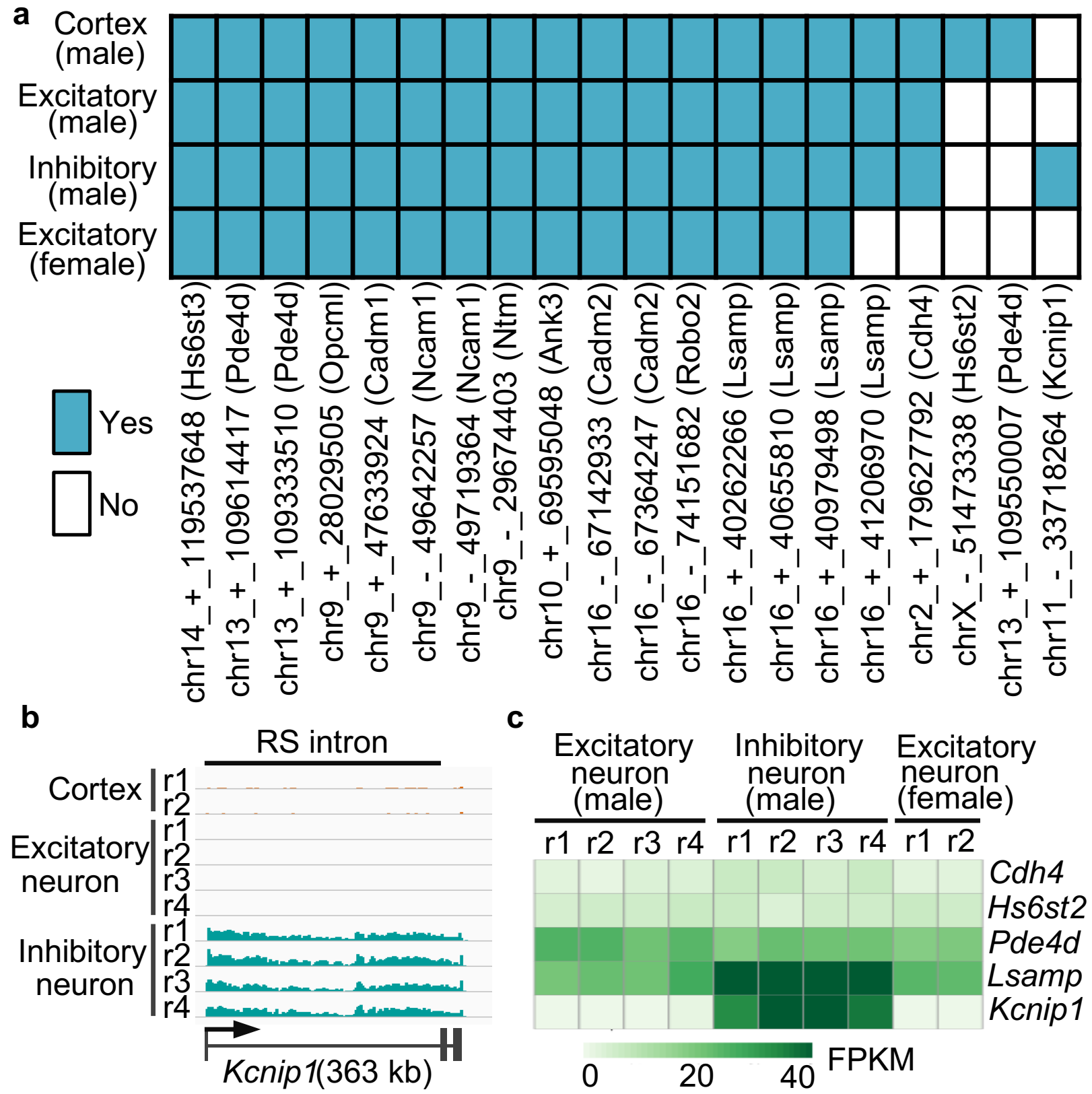
586 **Fig. 1** Nuclear total RNA-seq is efficient to identify recursive splicing (RS) events. **(a)**
587 Schematic of the two features of RS - the saw-tooth pattern (red triangles) and the RS
588 junction read. **(b)** Schematic of the isolation of different types of RNA. **(c)** Sequencing
589 profile at *Hs6st3* locus. r1, replicate 1. kb, kilobases. **(d)** Boxplot of normalized numbers
590 of RS junction reads at *Hs6st3* RS site. RPM, reads per million uniquely mapped reads.

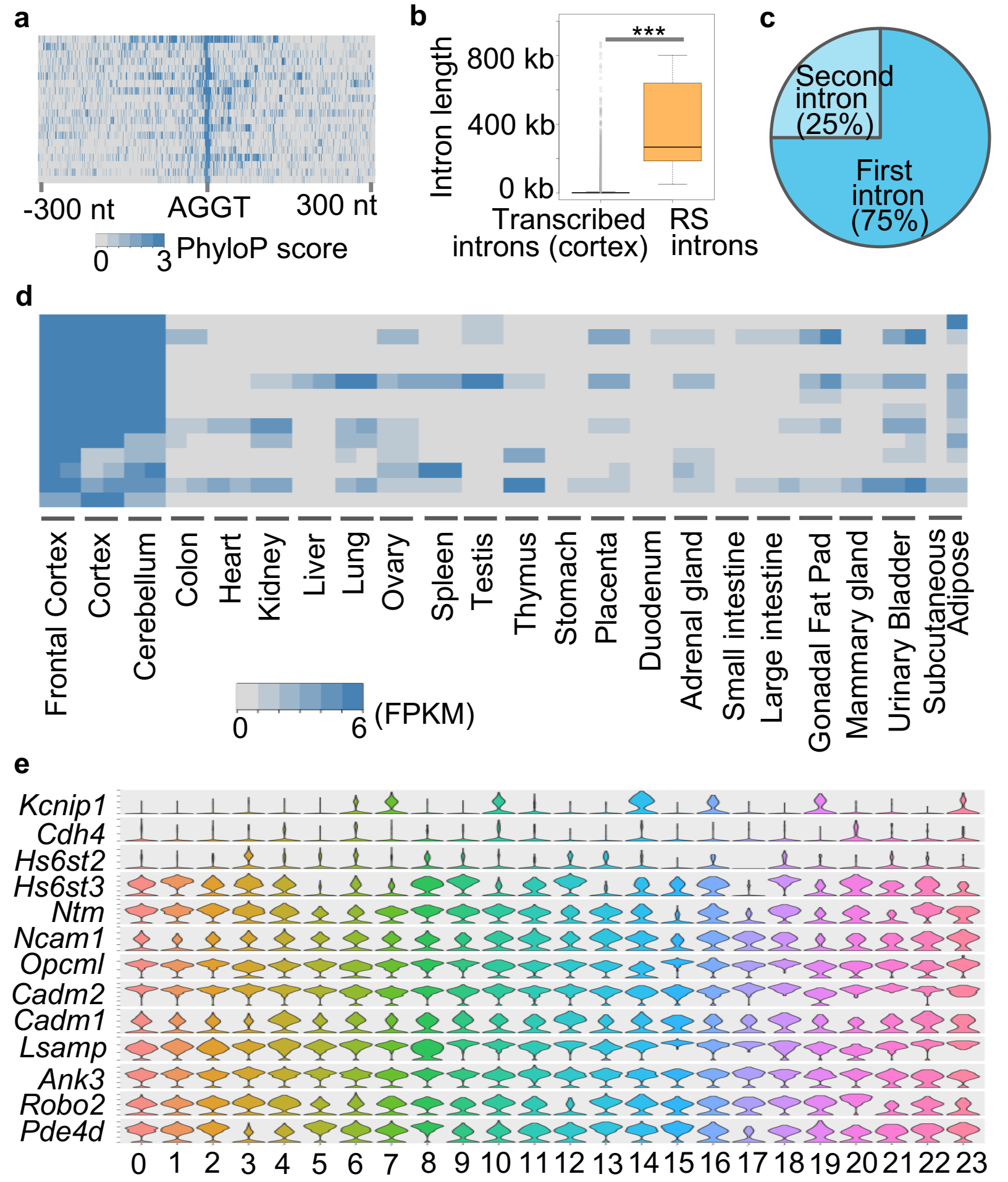
591 *, *P* = 0.03, one-tailed t-test. **(e)** Schematic of the pipeline utilizing nuclear total RNA-
592 seq data to identify RS sites. **(f)** Sequencing profile at *Lsamp* locus. Green arrows
593 indicate the four novel RS sites. **(g)** Bar plot of RS sites identified utilizing the two
594 sequencing 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

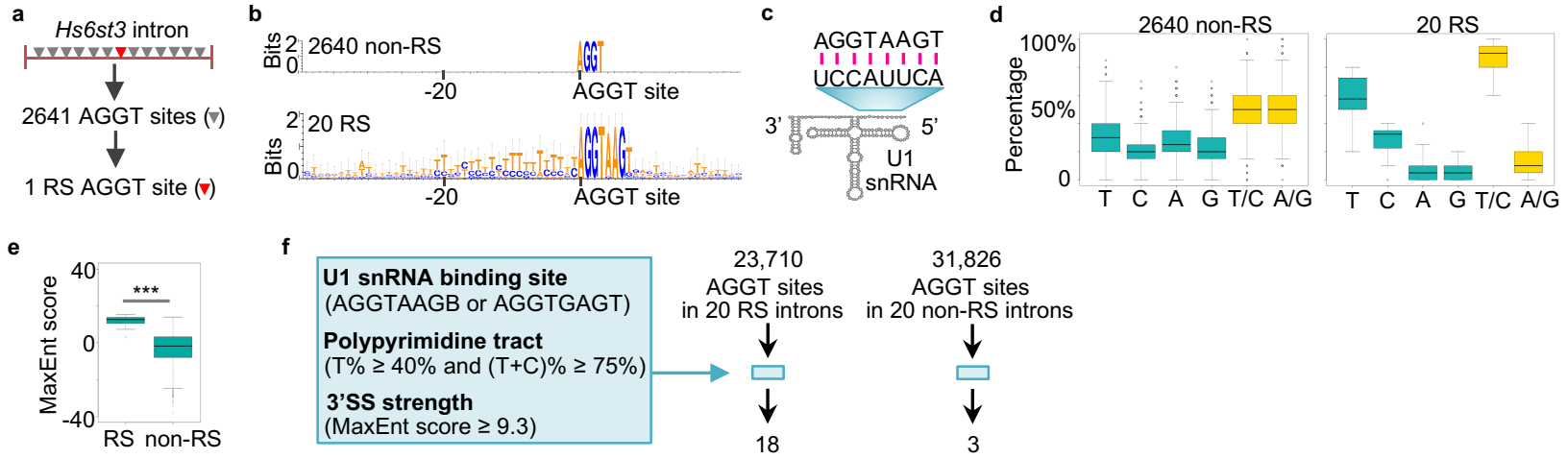
614 snRNA. **(d)** Boxplots of the percentages of nucleotides in the 20 nt region upstream of
615 the 2640 non-RS AGGT sites and the 20 RS AGGT sites. **(e)** Boxplot of MaxEnt 3'
616 splice site (3'SS) scores of the 20 RS AGGT sites and the 2640 non-RS AGGT sites.
617 ***, $P = 1.42^{-14}$, one-tailed t-test. **(f)** A computational pipeline to predict RS from intronic
618 AGGT sites. The criteria are listed in the blue box (left). B represents T, G, or C.
619
620 **Fig. 5** The exonicRS may represent the intermediate stage of RS evolving into
621 annotated RS exons. **(a)** Schematic of RS sites (RS) evolving into RS exons via the
622 exon-like RS events (exonicRS). Up, the junction reads spanning upstream exon;
623 Down, the junction reads spanning downstream exon. **(b)** Boxplots of the numbers of
624 the Up and Down junction reads of RS and exonicRS. ***, $P < 0.0001$, one-tailed t-test.
625 *ns*, $P = 0.29$, one-tailed t-test. **(c)** Sequence logos of the reconstituted 5'SS (r5'SS) and
626 the downstream 5'SS (Down 5'SS) of RS and exonicRS. **(d)** Boxplot of MaxEnt scores
627 of r5'SS and Down 5'SS of RS and exonicRS. ****, $P < 7.9^{-6}$, one-tailed t-test. **(e)**
628 Scatterplot of the MaxEnt scores of r5'SS and Down 5'SS of RS and exonicRS. **(f)**
629 Model for the classification of RS events based on the strengths of r5'SS and Down
630 5'SS.
631
632 **Supplemental Information**
633 **Additional file 1: Figure S1**. Novel RS sites. (**a**) The mapping statistics and access
634 numbers of RNA-seq data utilized in this study. (**b**) Pie charts of loci of uniquely mapped
635 reads in gene regions. (**c**) Schematic of the pipeline utilizing nuclear total RNA-seq data
636 to identify RS sites. (**d**) Heatmap of numbers of RS junction reads at each RS site in

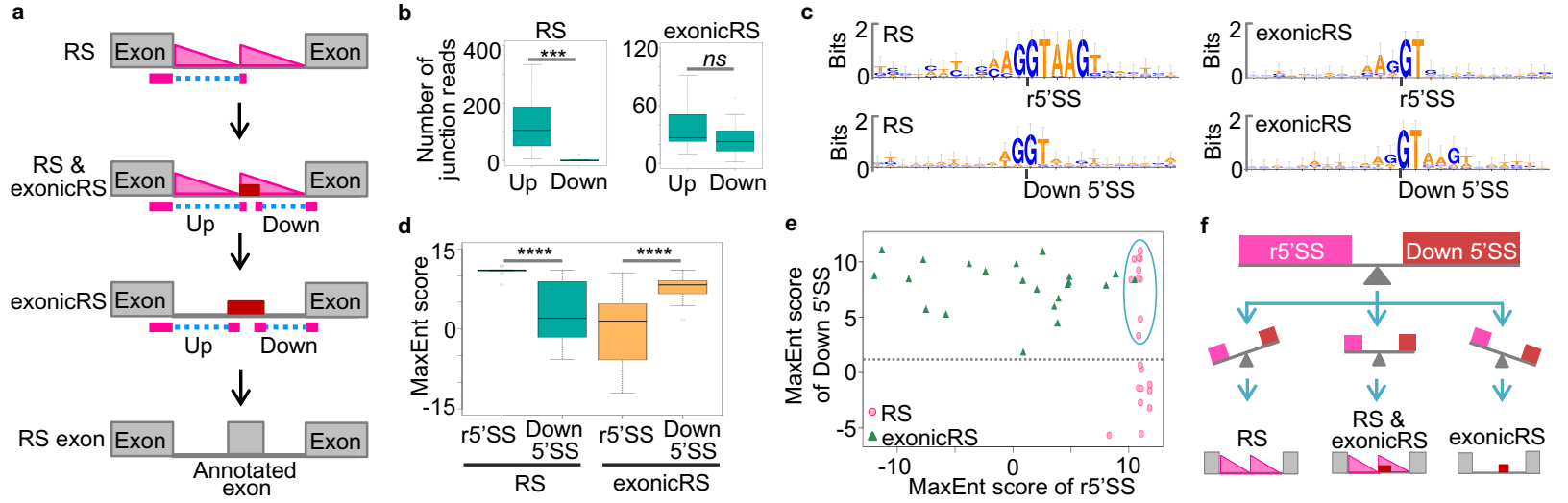
637 different total RNA-seq data sets. A green box indicates that the RS site was identified in
638 that data set. (**e**) Boxplots of the gene lengths (up) and expression levels (down) of genes
639 transcribed in the mouse cerebral cortex (transcribed gene) and RS genes. *P* indicates *P*
640 value, 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
RS intron
Exon
Exon
AGGT
RS junction read
b
mRNA
Whole-cell total RNA
Nuclear total RNA
c
RS intron
mRNA-seq
r1
r2
Whole-cell total RNA-seq
r1
r2
Nuclear total RNA-seq
r1
r2
Hs6st3 (731 kb)
d
RPM of RS junction reads
1
0.5
0
*
mRNA-seq
Whole-cell total RNA-seq
Nuclear total RNA-seq
e
Nuclear total RNA-seq
RS junction reads
AGGT
Intron profiles
RS sites
f
RS introns
mRNA-seq
r1
r2
Whole-cell total RNA-seq
r1
r2
Nuclear total RNA-seq
r1
r2
Lsamp (2197 kb)
g
Number of RS sites
0
10
20
Nuclear total RNA-seq
Whole-cell total RNA-seq


a
Cortex (male)
Excitatory (male)
Inhibitory (male)
Excitatory (female)
Yes
No
chr14_+_119537648 (Hs6st3)
chr13_+_109614417 (Pde4d)
chr13_+_109333510 (Pde4d)
chr9_+_28029505 (Opcml)
chr9_+_47633924 (Cadm1)
chr9_-_49642257 (Ncam1)
chr9_-_49719364 (Ncam1)
chr9_-_29674403 (Ntm)
chr10_+_69595048 (Ank3)
chr16_-_67142933 (Cadm2)
chr16_-_67364247 (Cadm2)
chr16_-_74151682 (Robo2)
chr16_+_40262266 (Lsamp)
chr16_+_40655810 (Lsamp)
chr16_+_40979498 (Lsamp)
chr16_+_41206970 (Lsamp)
chr2_+_179627792 (Cdh4)
chrX_-_51473338 (Hs6st2)
chr13_+_109550007 (Pde4d)
chr11_-_33718264 (Kcnip1)
b
RS intron
Cortex
r1
r2
Excitatory neuron
r1
r2
r3
r4
Inhibitory neuron
r1
r2
r3
r4
Kcnip1(363 kb)
c
Excitatory neuron (male)
Inhibitory neuron (male)
Excitatory neuron (female)
r1 r2 r3 r4 r1 r2 r3 r4 r1 r2
Cdh4
Hs6st2
Pde4d
Lsamp
Kcnip1
0
20
40
FPKM

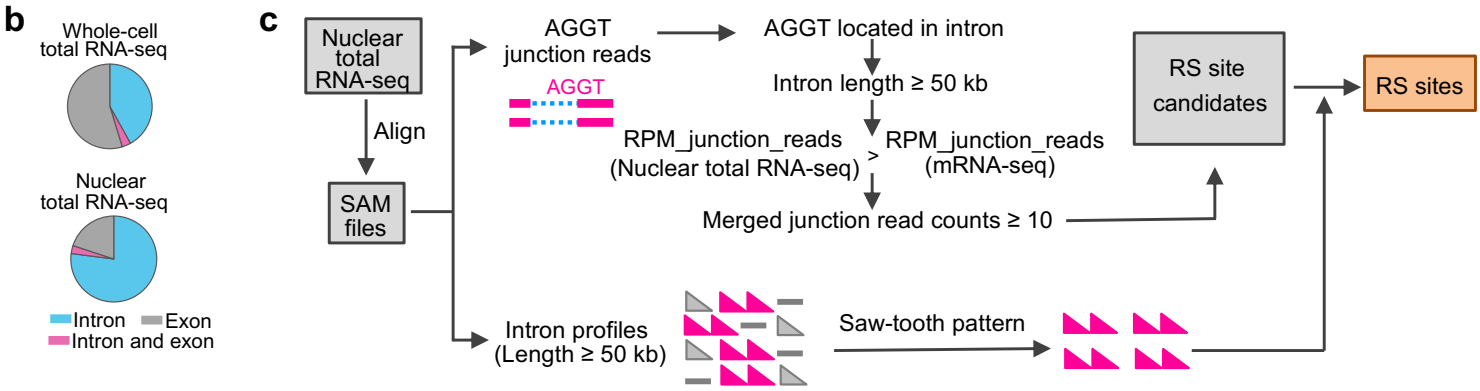

a
-300 nt
AGGT
300 nt
0
3
PhyloP score
b
Intron length
800 kb
400 kb
0 kb
***
Transcribed introns (cortex)
RS introns
c
Second intron (25%)
First intron (75%)
d
Frontal Cortex
Cortex
Cerebellum
Colon
Heart
Kidney
Liver
Lung
Ovary
Spleen
Testis
Thymus
Stomach
Placenta
Duodenum
Adrenal gland
Small intestine
Large intestine
Gonadal Fat Pad
Mammary gland
Urinary Bladder
Subcutaneous Adipose
0
6
(FPKM)
e
Kcnip1
Cdh4
Hs6st2
Hs6st3
Ntm
Ncam1
Opcml
Cadm2
Cadm1
Lsamp
Ank3
Robo2
Pde4d
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


a
Hs6st3 intron
2641 AGGT sites
1 RS AGGT site
b
Bits
2640 non-RS
-20
AGGT site
20 RS
c
AGGTAAGT
UCCAUUCA
3'
5'
U1
snRNA
d
2640 non-RS
20 RS
Percentage
100%
50%
0
T C A G T/C A/G
e
MaxEnt score
40
0
-40
***
RS non-RS
f
U1 snRNA binding site
(AGGTAAGB or AGGTGAGT)
Polypyrimidine tract
(T% ≥ 40% and (T+C)% ≥ 75%)
3'SS strength
(MaxEnt score ≥ 9.3)
23,710
AGGT sites
in 20 RS introns
18
31,826
AGGT sites
in 20 non-RS introns
3


a
RS
Exon
Exon
RS &
exonicRS
Exon
Exon
Up
Down
exonicRS
Exon
Exon
Up
Down
RS exon
Exon
Exon
Annotated
exon
b
RS
exonicRS
Number of
junction reads
400
200
0
***
120
60
0
ns
Up
Down
Up
Down
c
Bits
2
0
RS
r5'SS
Down 5'SS
exonicRS
d
MaxEnt score
15
0
-15
****
****
r5'SS
Down
5'SS
r5'SS
Down
5'SS
RS
exonicRS
e
MaxEnt score
of Down 5'SS
10
5
0
-5
RS
exonicRS
-10
0
10
MaxEnt score of r5'SS
f
r5'SS
Down 5'SS
RS
RS &
exonicRS
exonicRS

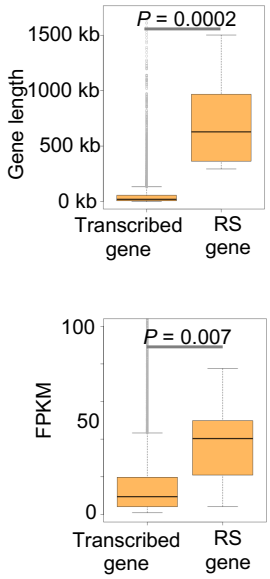
**a**

| | 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 |




**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 |

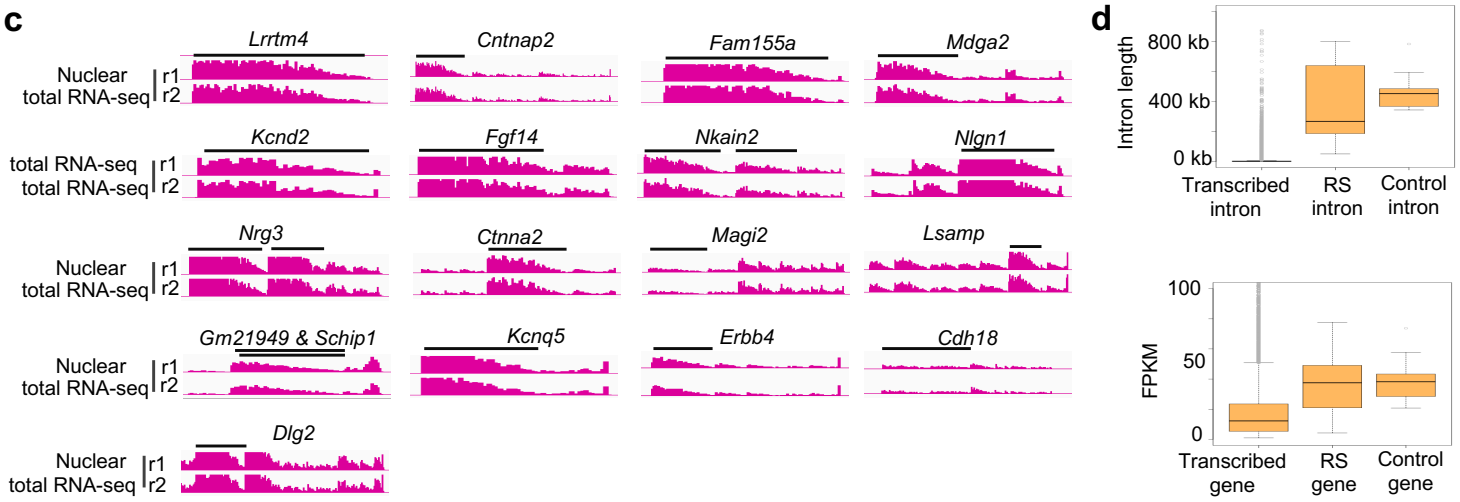
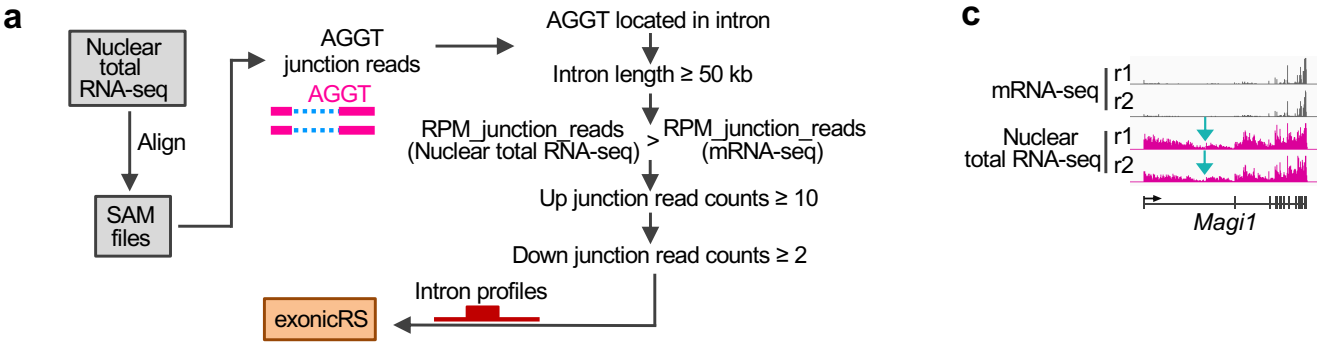



**Figure S2:** Primary sequence context of RS sites. (**a**) The sequence motifs, nucleotide percentages, and 3'SS MaxEnt scores of the 20 RS sites. The top 17 RS sites passed our prediction criteria of RS sites, while the bottom three RS sites failed. (**b**) The information of the non-RS AGGTs that 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.




**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.