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

Sohyun Moon[1], Ying-Tao 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

Recursive splicing (RS) is a splicing mechanism to remove long introns from messenger RNA precursors. RS is untraceable in the mature messenger RNA. Therefore, the RS splicing intermediates provide the direct evidence for the presence of RS events. However, RS splicing intermediates are transient and rapidly turned over, making them difficult to be captured. Here we demonstrate that nuclear RNA is enriched for RS splicing intermediates and nuclear total RNA sequencing is an efficient approach to investigate RS events. We analyzed high-depth nuclear total RNA sequencing data and identified 20 RS sites, suggesting that RS is a rare process of intron splicing in the mouse brain. We also identified that RS is constitutive between excitatory and inhibitory neurons and between sexes in the mouse cerebral cortex. In addition, we found that the primary sequence context is associated with RS splicing intermediates and distinguishes RS AGGT site from non-RS AGGT sites, indicating the importance of the primary sequence context in RS sites. Moreover, we discovered that cryptic exons may use an RS-like mechanism for splicing. Overall, our findings provide novel insights into the mechanisms of RS in long genes.

47

48 **Introduction**

49 Recursive splicing (RS) is a splicing mechanism that removes a long intron into
50 several smaller segments as opposed to in a large single unit[1-11]. RS is untraceable in the
51 mature messenger RNA (mRNA), and the direct evidence of RS is the splicing
52 intermediates. However, RS splicing intermediates are unstable, making them difficult to
53 be captured and analyzed. Whole-cell ribosomal RNA-depleted total RNA sequencing
54 (total RNA-seq) and nascent RNA-seq have been used to identify RS splicing
55 intermediates[3,5,7-9,11], but their efficiency is relatively low. Thus, new approaches are
56 needed to investigate RS events.

57 RS has been widely studied in mammals. Two earlier studies identified five and nine
58 RS events in humans and ten RS events in mice[3,5], suggesting that RS might be a rare
59 splicing mechanism for a small group of long introns. Intriguingly, a recent study using
60 nascent RNA-seq identified 342 candidate RS sites in three human cell lines[9]. Similarly,
61 another study using nascent RNA-seq reported 5,468 RS events in a human cell line[11].
62 These two studies suggested that RS might be a widely used splicing mechanism in
63 mammals. Given the inconsistency, further studies are required to determine the extent
64 to which RS is used for intron splicing in mammals.

65 Here, we report that nuclear total RNA-seq is enriched for RS splicing intermediates
66 and nascent transcripts, which suggests that it is an efficient approach to investigate RS
67 events. Using nuclear total RNA-seq data generated from the mouse brain, we identified
68 novel RS events, examined the cell-type and sex specificity of RS, and analyzed RS
69 splicing intermediates. We found that RS is a rare process of intron splicing in the mouse

brain. We also found that some cryptic exons use an RS-like mechanism for splicing. Together, our findings provide novel insights into the mechanisms of RS in long genes.

Results

Nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates.

To determine the extent to which nascent transcripts and RS splicing intermediates are enriched in nuclear RNA (Fig. 1a), we analyzed nuclear total RNA-seq and whole-cell total RNA-seq data that we generated from the mouse cerebral cortex[12,13] (Supplementary Figure S1a). As a control data set, we also analyzed poly(A) enriched messenger RNA-seq (mRNA-seq) data that were generated from the same mouse brain region[14].

Nascent transcripts, which can be indicated by the sequencing reads mapped to introns[15], are important to identify RS events, because RS introns exhibit a saw-tooth pattern of nascent transcript signals from total RNA-seq data[3,5] (red triangles, Fig. 1b). To determine the enrichment of nascent transcripts in nuclear RNA, we calculated the proportions of reads mapped to introns in nuclear total RNA-seq data, whole-cell

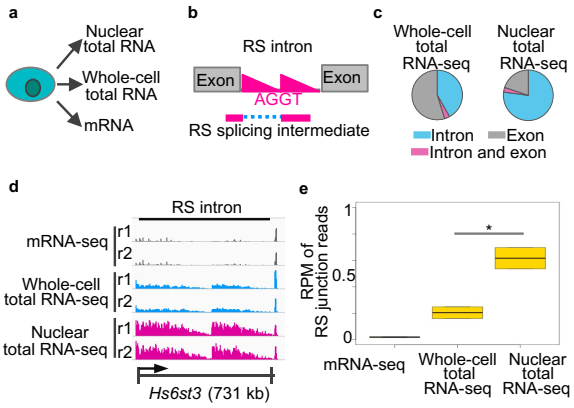


Fig. 1 Nuclear total RNA is enriched for nascent transcripts and RS splicing intermediates. **(a)** Schematic of the isolation of different types of RNA. **(b)** Schematic of the two features of RS, the saw-tooth pattern (red triangles) and the RS splicing intermediate. **(c)** Pie charts of loci of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

93 total RNA-seq data, and mRNA-seq data. We found that 77% of the uniquely mapped
94 reads from the nuclear total RNA-seq data are localized in introns, which is significantly
95 higher than the 41% from the whole-cell total RNA-seq data and the 23% from the mRNA-
96 seq data ($P < 2.2^{-16}$, one-tailed Fisher's Exact Test) (Fig. 1c and Supplementary Figure
97 S1a). In addition, at a known RS intron in *Hs6st3*, we observed a more distinct saw-tooth
98 pattern in the nuclear total RNA-seq data than in the whole-cell total RNA-seq data (Fig.
99 1d). These results suggest that nuclear total RNA is enriched for nascent transcripts.
100 RS splicing intermediates, which can be indicated by the junction reads spanning
101 the upstream exon and the RS AGGT site (RS junction reads, Fig. 1b), is also important
102 to identify RS events. To examine RS splicing intermediates in nuclear RNA, we
103 calculated the numbers of junction reads at a known RS site in *Hs6st3*. We found that the
104 number of RS junction reads at *Hs6st3* is 3-fold higher in nuclear total RNA-seq data than
105 in whole-cell total RNA-seq data (Fig. 1e). Together, these results indicate that nuclear
106 total RNA is enriched for nascent transcripts and RS splicing intermediates.
107
108 **Identify RS sites using nuclear total RNA-seq data.**
109 To identify RS sites from nuclear total RNA-seq data, we modified a previously developed
110 pipeline and incorporated additional criteria[5] (Fig. 2a). First, we extracted all junction
111 reads spanning the AGGT sites in the mouse genome and then focused on the AGGT
112 sites located in introns that are longer than 1 kb. To exclude unannotated exons from our
113 analyses, we selected AGGT sites that show more than two-fold enrichment of RS
114 junction reads in nuclear total RNA-seq than in mRNA-seq. To increase the degree of
115 confidence, we selected sites containing 10 or more RS junction reads as RS site

116 candidates. To enhance the detection of RS junction reads, one of the nuclear total RNA-
117 seq sample was sequenced to extra depth, resulting in 428 million uniquely mapped reads
118 (Supplementary Figure S1a). Lastly, we visually inspected the candidates to select sites
119 that show saw-tooth patterns in their host introns as RS sites.

120 By applying this pipeline to our nuclear total RNA-seq data, we identified 84 RS site
121 candidates (Supplementary Figure S1b) and 19 RS sites (Supplementary Figure S1c).
122 The 19 RS sites include all the ten known RS sites in mice[5] (Supplementary Figure S1c).
123 Notably, nine RS sites (47%) we identified are novel in mice, including the four sites in
124 the introns of *Lsamp* (Fig. 2b, green arrows). To validate the identified RS sites, we
125 conducted RT-PCR to detect the RS splicing intermediates for three known RS sites in
126 *Hs6st3* and *Cadm2* and three novel RS sites in *Lsamp* (Fig. 2c). Notably, we detected RS
127 splicing intermediates for all the six RS sites (Fig. 2c), which independently supported the
128 RS sites that we identified using the nuclear total RNA-seq data.

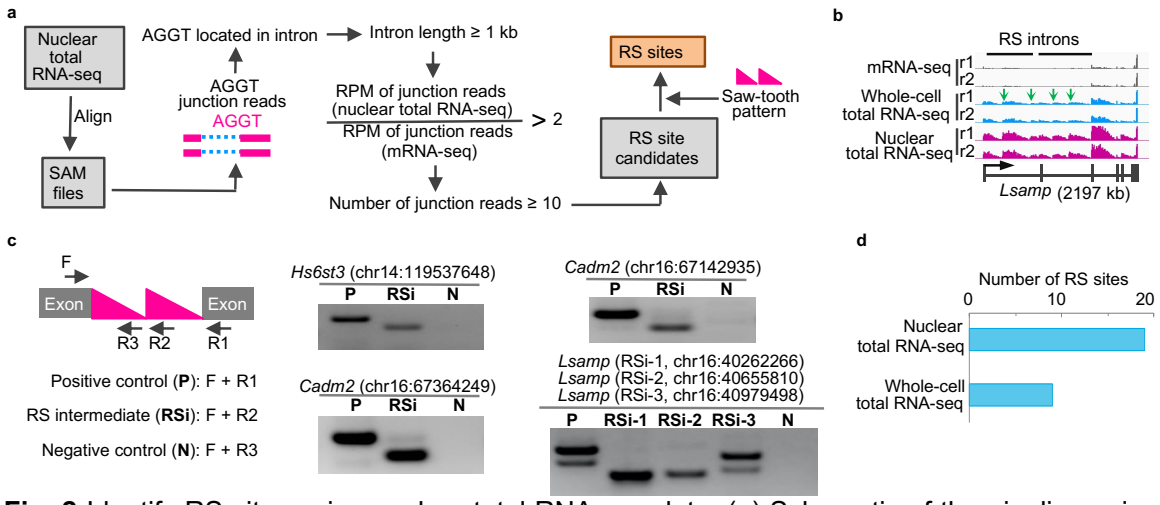


Fig. 2 Identify RS sites using nuclear total RNA-seq data. (**a**) Schematic of the pipeline using nuclear total RNA-seq data to identify RS sites. (**b**) Sequencing profile at *Lsamp* locus. Green arrows indicate the four novel RS sites. **(c)** Using RT-PCR to detect RS splicing intermediates. Left,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

129 We next applied the same pipeline to the whole-cell total RNA-seq data and
130 identified nine RS sites (Supplementary Figure S1c). We found that using nuclear total
131 RNA-seq identified two-fold more RS sites than using whole-cell total RNA-seq (Fig. 2d),
132 suggesting that nuclear total RNA-seq is an efficient approach to identify RS events.
133 To determine whether our criteria were too stringent, we lowered the cutoff of the
134 counts of RS junction reads from 10 to five. We identified 103 additional RS site
135 candidates (Supplementary Figure S1d). However, none of them exhibited saw-tooth
136 patterns in the host introns, suggesting that lowering the cutoff of the counts of RS junction
137 reads did not identify additional RS sites.
138
139 **RS is restricted to the AGGT motif.**
140 Although GT is the most highly used motif for 5' splice sites (5'SS), other motifs can also
141 be used for 5'SS. To determine whether RS occurs in AGNN motifs where NN is not GT,
142 we used a pipeline to identify RS sites in AGNN motif. This pipeline is similar to the one
143 in Fig. 2a. Briefly, we extracted all junction reads spanning the AGNN sites, focused on
144 the AGNN sites located in introns, selected AGNN sites that show more than two-fold
145 enrichment of RS junction reads in nuclear total RNA-seq than in mRNA-seq, and then
146 selected sites containing 10 or more RS junction reads as RS site candidates. As a result,
147 we identified 405 AGNN RS site candidates (Supplementary Figure S2). However, visual
148 inspection revealed that none of them exhibit saw-tooth patterns in the host introns. Thus,
149 these results indicate that RS is restricted to the AGGT motif.
150
151

Cell-type and sex specificity of RS in the mouse cerebral cortex.

To determine the cell-type specificity of RS, we focused on two types of cortical neurons, the excitatory neurons and the inhibitory neurons, which account for 85% and 15% of the neurons in the mouse cerebral cortex[12]. We analyzed the nuclear total RNA-seq data that we generated from the two neuronal cell types (Supplementary Figure S1a). We applied our pipeline to these data and identified 17 RS sites in excitatory neurons and 18 RS sites in inhibitory neurons (Fig. 3a and Supplementary Figure S1c). Notably, all but one of the RS sites are common in both neuronal cell types (Fig. 3a). The only RS site unique to inhibitory neurons resides in the *Kcnip1* gene, which is only expressed in the inhibitory neurons (Fig. 3b and 3c). Thus, these results indicate that RS is largely constitutive between the two types of neurons in the mouse cerebral cortex.

To investigate the sex specificity of RS, we analyzed the nuclear total RNA-seq data of excitatory neurons from the cerebral cortex of female mice[12]. We identified 15 RS sites (Supplementary Figure S1c), which are all included in the 17 RS sites we identified from male excitatory neurons (Fig. 3a). The remaining two RS sites are unlikely male specific, because we also identified seven and

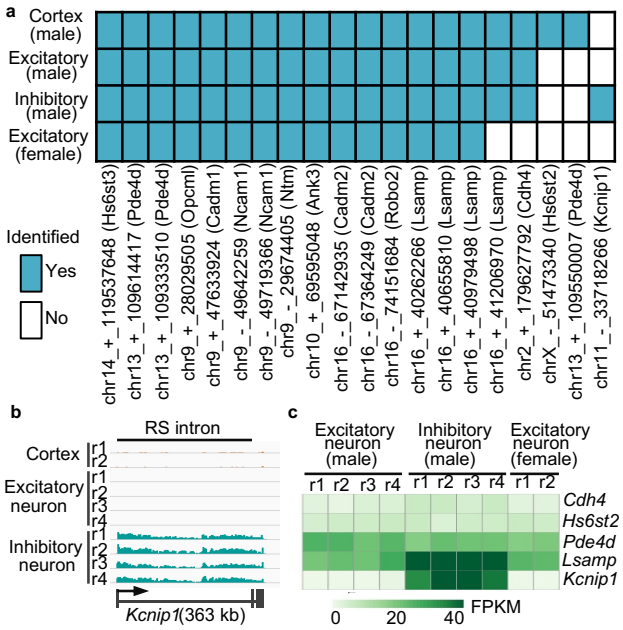


Fig. 3 Cell-type and sex specificity of RS in the mouse cortex. **(a)** Heatmap of RS sites identified in male cortex, male cortical excitatory and inhibitory neurons, and female cortical excitatory neurons. **(b)** Nuclear total RNA-seq profile (male) at *Kcnip1* locus. **(c)** Heatmap of expression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

175 five junction reads for them in the female data (Supplementary Figure S1c), although they
176 failed to pass our criteria of 10 junction reads. Together, these results indicate that RS is
177 constitutive between male and female excitatory neurons in the mouse cerebral cortex.
178
179 **Characteristics of RS sites.**
180 In consistent with previous reports, RS sites are highly conserved (Supplementary Figure
181 S3a), are located in long introns (Supplementary Figure S3b), prefer the first two introns
182 (Supplementary Figure S3c, 75% in first intron and 25% in second intron), and are located
183 in genes that are specifically expressed in the brain (Supplementary Figure S3d).
184 Furthermore, compared to non-RS AGGT sites, RS AGGT sites are enriched for
185 AGGTAAGT motif that complements with the 5' conserved sequence of U1 snRNA
186 (Supplementary Figure S3e-g), are enriched for thymine and cytosine in the 20 nt regions
187 upstream of RS AGGT sites (polypyrimidine tract, Supplementary Figure S3h), and show
188 higher 3' splice site scores when examined using MaxEntScan[16] (Supplementary Figure
189 S3i).
190
191 **Splicing intermediates linking RS exons to the downstream annotated exons.**
192 RS uses exon definition mechanism[5,7], thus suggesting a possibility of splicing
193 intermediates linking RS exons to the downstream annotated exons (downstream splicing
194 intermediates, Fig. 4a). Although Sibley et al. reported that RS exons were not detectable
195 in mRNA transcripts and were included only after blocking the 5'SS using antisense
196 oligonucleotide[5], the extent to which the downstream splicing intermediates exist in
197 physiological conditions remains largely unexplored.

198 To identify downstream splicing intermediates of RS sites, we developed a
199 computational pipeline to identify junction reads that link RS exons to downstream
200 annotated exons (down-junction reads, Fig. 4a). Briefly, we obtained all the junction reads
201 from our nuclear total RNA-seq data and then scanned the 350 nt region downstream of
202 the RS AGGT site for junction reads that link to the downstream exons. We found down-
203 junction reads for 10 RS sites, but not for the other 10 sites (Supplementary Figure S4a).
204 To validate the findings of down-junction reads, we conducted RT-PCR to detect the
205 downstream splicing intermediates for four RS sites (Fig. 4b). Notably, for the two RS
206 sites that show both up- and down-junction reads, we detected both upstream and
207 downstream splicing intermediates (Fig. 4b, left panel). However, for the two RS sites that
208 only show up-junction reads, we only detected the upstream splicing intermediates, not
209 the downstream splicing intermediates (Fig. 4b, right panel). Thus, the RT-PCR results
210 are consistent with the results from the nuclear total RNA-seq. Together, these results

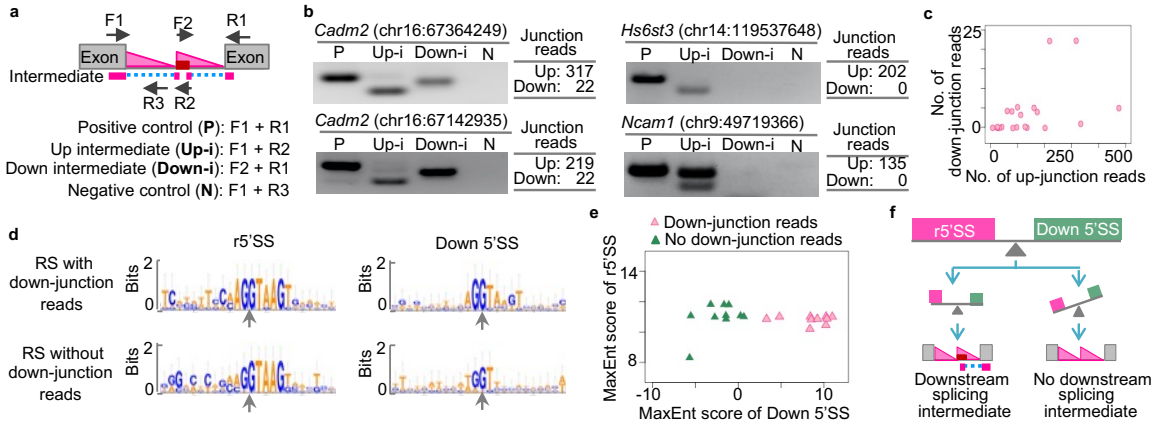


Fig. 4 Downstream RS splicing intermediates. **(a)** Schematic of the up- and down-junction reads that represent the upstream and downstream RS splicing intermediates and the primer design. **(b)** Validation of the upstream (Up-i) and downstream (Down-i) RS splicing intermediates using RT-PCR. P, positive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

211 indicate that ten RS sites have downstream splicing intermediates, while the other ten
212 sites do not have downstream splicing intermediates.
213
214 **Downstream splicing intermediates are associated with high strengths of the 5'**
215 **splice sites at the 3' end of the RS exons.**
216 To understand why some RS sites do not show downstream splicing intermediates, we
217 first asked whether these RS sites also show low abundance of upstream splicing
218 intermediates. We analyzed the numbers of up- and down-junction reads for the 20 RS
219 sites. However, we found no positive correlation between the two (Fig. 4c), suggesting
220 that the absence of downstream splicing intermediates for the 10 RS sites is not
221 associated with the low abundance of upstream splicing intermediates.
222 Previous studies have showed that the competition between the two 5'SS was
223 associated with the inclusion of the RS exons in mature transcripts[5,7,9]. Thus, another
224 hypothesis is that the 5'SS at the two ends of the RS exons, the reconstituted 5'SS (r5'SS)
225 after the first step of splicing and the downstream 5'SS (Down 5'SS), affect the existence
226 of the downstream splicing intermediates. To test this hypothesis, we quantified the
227 strengths of r5'SS and Down 5'SS using MaxEntScan[16]. We found that RS sites with
228 down-junction reads show higher MaxEnt scores of Down 5'SS compared to RS sites
229 without down-junction reads ($P = 4.28^{-9}$, one-tailed t-test) (Fig. 4d,e and Supplementary
230 Figure S4a). By contrast, the scores of r5'SS are comparable between the two groups
231 (Fig. 4d,e and Supplementary Figure S4a). Together, these results support a model that
232 the downstream splicing intermediates of RS sites are likely associated with the high
233 strengths of 5'SS at the 3' end of the RS exons (Fig. 4f).

234

235 **Cryptic exons use a RS-like splicing mechanism.**

236 During our analysis of RS sites, we found that some cryptic exons, although showing no
237 saw-tooth patterns in the host introns, exhibit RS-like features (hereafter termed RS-like
238 cryptic exons, Fig. 5a). To systematically identify RS-like cryptic exons, we developed a
239 computational pipeline (Supplementary Figure S4b). To exclude unannotated exons from
240 our downstream analysis, we used a criterion that the normalized count of up-junction
241 reads in nuclear total RNA-seq should
242 be higher than that in mRNA-seq. By
243 applying this pipeline to our nuclear
244 total RNA-seq data, we identified 22
245 RS-like cryptic exons in the introns of
246 21 long genes (Supplementary Figure
247 S4c). To validate the RS-like cryptic
248 exons, we conducted RT-PCR to
249 detect splicing intermediates for three
250 RS-like cryptic exons. We confirmed
251 the upstream and downstream splicing
252 intermediates for all the three RS-like
253 cryptic exons (Fig. 5b). Notably, we
254 found that although the cryptic exons
255 in *Lrrc4c* and *Magi1* were not included
256 in the mature mRNA, the cryptic exon

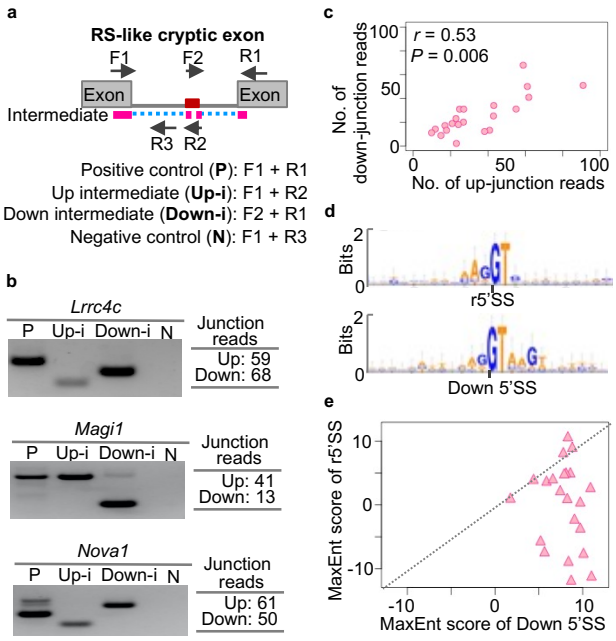


Fig. 5 Cryptic exons use a RS-like splicing mechanism. **(a)** Schematic of RS-like cryptic exons and the primer design. **(b)** Validation of the upstream (Up-i) and downstream (Down-i) splicing intermediates using RT-PCR. P, positive control. N, negative control. Left, RT-PCR and gel results. Right,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

257 in *Nova1* was included in a part of the mature mRNA (Fig. 5b), suggesting that RS-like
258 cryptic exons may be included in the mature mRNA.

259 To identify features of RS-like cryptic exons, we analyzed the correlation between
260 the up- and down-junction reads for them. We found that RS-like cryptic exons exhibit a
261 positive correlation between the numbers of up- and down-junction reads (Pearson
262 correlation coefficient r = 0.53, *P-value* = 0.01, Fig. 5c), which contrasts with the results
263 of RS sites (Fig. 4c). To investigate the strengths of the 5'SS at the two ends of RS-like
264 cryptic exons, we examined the sequence compositions of the r5'SS and Down 5'SS
265 using WebLogo[17]. We found that the RS-like cryptic exons show an enrichment of 5'SS
266 motif (AGGTAAGT) at the Down 5'SS but not at the r5'SS (Fig. 5d). Furthermore, we
267 quantified the strengths of r5'SS and Down 5'SS using MaxEntScan[16] and found that the
268 MaxEnt scores of r5'SS are significantly lower than that of Down 5'SS in RS-like cryptic
269 exons (*P-value* < 0.0001, one-tailed t-test, Fig. 5e). One exception is the RS-like cryptic
270 exon in *Magi1*, which shows a much higher MaxEnt score at the r5'SS than the Down
271 5'SS (Fig. 5e and Supplementary Figure S4c). Notably, the *Magi1* intron exhibits a weak
272 saw-tooth pattern (Supplementary Figure S4d), which failed to pass the stringent criteria
273 in our identification of RS sites. Taken together, these results demonstrate that some
274 cryptic exons may use an RS-like mechanism for splicing.

275

276

277

278

279

280 **Discussion**

281 In this study, we demonstrated that nuclear total RNA-seq is an efficient approach
282 to investigate RS events. We developed a novel pipeline to identify RS sites, determined
283 the cell-type and sex specificity of RS, and characterized the genomic features of RS
284 sites. Through analysis of the primary sequences, we found that RS splicing
285 intermediates are associated with the primary sequence context. We also discovered that
286 some cryptic exons use an RS-like mechanism for splicing.

287 The extent to which RS is used for intron splicing in mammals remains inconclusive.
288 Here we analyzed high-depth nuclear total RNA-seq data and identified 20 RS sites in
289 mice, which is consistent with previous studies[3,5] that RS is a rare splicing process for a
290 small group of long introns in the mammalian genome. We found that whether or not
291 including the saw-tooth patterns to define RS sites might be the major reason for the
292 differences of the reported numbers of RS events[3,5,9,11]. For example, we identified 65
293 AGGT sites and 405 AGNN sites that have 10 or more junction reads, but they did not
294 show saw-tooth patterns in the host introns. Therefore, we did not consider them as RS
295 sites. Notably, the relationship between saw-tooth patterns and RS events remains
296 unclear. Further studies are needed to investigate whether the saw-tooth pattern should
297 be a required criterion to define an RS event.

298 Long introns exhibit a high rate of creating new exons during evolution[18], but the
299 underlying mechanisms are not fully understood. Our discovery of RS-like cryptic exons
300 indicates that long introns may acquire novel exons via the RS-like cryptic exons. This is
301 supported by the findings that more than 6000 human annotated exons are RS-like
302 exons[19]. In addition, the numbers of RS sites in *Drosophila melanogaster* are about 15

303 times more than that in humans[3,5,7-9], but the numbers of RS-like annotated exons in
304 *Drosophila* are 2~100 times less than that in humans[5,7,19]. Thus, future studies are
305 needed to investigate RS, RS-like cryptic exons, and RS-like annotated exons in
306 evolutionarily distinct species to determine their associations during evolution.

307 RS genes are genetically linked to various human brain disorders. For example,
308 *ANK3* encodes ankyrin-G and is linked to autism spectrum disorders, attention deficit
309 hyperactivity disorder, intellectual disability, and bipolar disorder[20-23]. Also, *NTM* encodes
310 neurotrimin and is linked to autism spectrum disorders and attention deficit hyperactivity
311 disorder[24,25]. The *PDE4D* encodes phosphodiesterase 4D and is linked to schizophrenia,
312 psychosis, acrodysostosis, and neuroticism[26-28]. Notably, PDE4D Inhibitors are in clinical
313 trials for the treatment of Alzheimer's disease and Fragile X syndrome[29,30]. Given that the
314 disruption of the RS process interfered with the RS gene function and caused abnormality
315 in the central nervous system[7], further investigation will be necessary to illuminate
316 whether RS mechanism contributes to the pathophysiology of these human brain
317 disorders.

318

319

320 **Methods**

321 **Animals.** The C57BL/6 mice were obtained from the Jackson Laboratory. Mice were
322 maintained in the same genetic background and were kept on a regular 12-hour light/12-
323 hour dark cycle. Animal studies were carried out in accordance with the National Institutes
324 of Health's Guide for the Care and Use of Laboratory Animals recommendations. The
325 study protocol was approved by the Institutional Animal Care and Use Committee

326 (IACUC) of the New York Institute of Technology. The study is reported in accordance
327 with the ARRIVE guidelines (https://arriveguidelines.org).
328
329 **RNA isolation and RT-PCR.**
330 The brain was dissected from 8-week-old mice. Total RNA was extracted from the brain
331 tissues using TRIzol reagent (Invitrogen, #15596026) according to the manufacturer's
332 instructions. The integrity and quality of the isolated RNA were determined by the
333 Nanodrop and the Bioanalyzer. The Prime script reverse transcript kit (Takara, #RR037A)
334 was used to synthesize the cDNA. The RT-PCR experiments were conducted using the
335 GoTaq G2 Hot Start Master mix (Promega #M7422). The PCR primers were listed in the
336 Supplementary Figure S5. The RT-PCR were conducted using the following setting: 94 °C
337 for 3 minutes, followed by 32-40 cycles of thermocycling (94 °C for 30 seconds, 52 °C-
338 56 °C for 30 seconds, 72 °C for 30 seconds), and 72 °C for 3 minutes.
339 **Statistical analysis.** All statistical analyses were performed in the R software version
340 3.6.1 (https://www.r-project.org/).
341 **Nuclear total RNA-seq data analysis.** The analysis pipeline was described
342 previously[12,13,31]. In brief, raw data in sra files were downloaded from the EBI European
343 Nucleotide Archive database[32] using the accession numbers listed in Figure S1. The
344 fastq-dump.2.9.6 of NCBI SRA ToolKit was used to extract the FASTQ files using the
345 parameter of "--split-3". STAR[33] was used to map the FASTQ raw reads into mouse mm10
346 genome using the parameters of "--runThreadN 40 --outFilterMultimapNmax 1 --
347 outFilterMismatchNmax 3". The samtools view[34] was used to convert sam files into bam
348 files. The samtools sort was used to sort the bam files. The samtools index was used to

349 index the sorted bam files. The bamCoverage[35] was used to convert the sorted bam files
350 into strand-specific bigwig files. The bamCoverage parameters that were used included
351 "--filterRNAstrand forward --binSize 1 -p 40 -o" for plus strand and "--filterRNAstrand
352 reverse --binSize 1 -p 40 -o" for minus strand.

353 **Sequencing data visualization.** All sequencing data, including the bigwig files and bam
354 files, were visualized in the IGV_2.8.2 genome browser[36].

355 **Genome annotation.** The gtf file of mouse genome annotation was downloaded from the
356 Ensembl release 93[37].

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

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

368 **Pipeline to identify RS sites.** The schematic of this pipeline is shown in Fig. 2a. Briefly,
369 AGGT sites located in introns longer than or equal to 1 kb were extracted. Sites that
370 showed a larger than two-fold RPM value of junction reads in total RNA-seq data than in
371 mRNA-seq data were kept for downstream analyses. The counts of junction reads of

372 biological replicates were merged, and AGGT sites that contained 10 or more junction
373 reads were identified as RS site candidates. The RS site candidates were further refined
374 as RS sites if the host intron showed a clear saw-tooth pattern by visual inspection.

375 **FPKM of nuclear total RNA-seq data.** The number of reads mapped to the exonic
376 regions of each gene were calculated to get the raw counts. The raw counts were then
377 normalized to the exon lengths of that gene and to the sequencing depth of that data set
378 to get the FPKM values.

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

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

386 **WebLogo analysis.** WebLogo 3[17] (http://weblogo.threeplusone.com) was used to
387 perform the sequence logo analysis. The Output Format was chosen as "PNG (high res.)",
388 and the Stacks per Line was set to "80". The default values were used for other
389 parameters.

390 **MaxEntScan 3'SS analysis.** The 3'SS scores were calculated by MaxEntScan::score3ss
391 (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html)[16]. The
392 input sequences were composed of the 18 nt region upstream of the AGGT, the AGGT
393 motif, and the one nucleotide following AGGT (18nt + AGGT + 1nt). The three models -

394 Maximum Entropy Model, First-order Markov Model, and Weight Matrix Model - were
395 selected. The MaxEnt scores were used as the 3'SS scores.
396 **Reconstituted 5' splice sites (r5'SS).** The r5'SS sequences were composed of the last
397 30 nucleotides of the upstream exon, the GT motif, and the 20 nucleotides following
398 AGGT (30nt + GT + 20nt).
399 **MaxEntScan 5'SS analysis.** The 5'SS scores were calculated by MaxEntScan::score5ss
400 (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)[16]. The input
401 sequences for 5'SS were composed of three nucleotides before the GT, the GT motif,
402 and the four nucleotides following AGGT (3nt + GT + 4nt). The input sequences of r5'SS
403 and Down 5'SS were listed in Supplementary Figure S4a and S4c.
404 **Pipeline to identify RS-like cryptic exons.** The schematic of this pipeline is shown in
405 Supplementary Figure S4b. Briefly, AGGT sites located in introns longer than or equal to
406 50 kb were extracted. The AGGT sites that showed a larger RPM value of junction reads
407 in total RNA-seq data than in mRNA-seq data were kept for downstream analyses. The
408 counts of the junction reads of the biological replicates were merged. AGGT sites
409 contained 10 or more up-junction reads and two or more down-junction reads were
410 identified as candidates of RS-like cryptic exons. The RS-like cryptic exon candidates
411 were further refined as RS-like cryptic exons if the host intron showed an exon-like but
412 not saw-tooth like pattern.
413
414 **Data availability**
415 The data sets supporting the conclusions of this article are available in the NCBI GEO
416 database with the accession numbers listed in Supplementary Figure S1a. 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.

Author contributions

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

Acknowledgements

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

References

1 Hatton, A. R., Subramaniam, V. & Lopez, A. J. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Mol Cell* **2**, 787-796, doi:10.1016/s1097-2765(00)80293-2 (1998).

2 Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J. & Lopez, A. J. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics* **170**, 661-674, doi:10.1534/genetics.104.039701 (2005).

3 Duff, M. O. *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521**, 376-379, doi:10.1038/nature14475 (2015).

4 Kelly, S. *et al.* Splicing of many human genes involves sites embedded within introns. *Nucleic Acids Res* **43**, 4721-4732, doi:10.1093/nar/gkv386 (2015).

5 Sibley, C. R. *et al.* Recursive splicing in long vertebrate genes. *Nature* **521**, 371-375, doi:10.1038/nature14466 (2015).

446 6 Hayashi, T. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics
447 of recursive splicing and enhancer RNAs. *Nat Commun* **9**, 619,
448 doi:10.1038/s41467-018-02866-0 (2018).
449 7 Joseph, B., Kondo, S. & Lai, E. C. Short cryptic exons mediate recursive splicing
450 in Drosophila. *Nat Struct Mol Biol* **25**, 365-371, doi:10.1038/s41594-018-0052-6
451 (2018).
452 8 Pai, A. A., Paggi, J. M., Yan, P., Adelman, K. & Burge, C. B. Numerous recursive
453 sites contribute to accuracy of splicing in long introns in flies. *PLoS Genet* **14**,
454 e1007588, doi:10.1371/journal.pgen.1007588 (2018).
455 9 Zhang, X. O., Fu, Y., Mou, H., Xue, W. & Weng, Z. The temporal landscape of
456 recursive splicing during Pol II transcription elongation in human cells. *PLoS Genet*
457 **14**, e1007579, doi:10.1371/journal.pgen.1007579 (2018).
458 10 Joseph, B. & Lai, E. C. The exon junction complex and intron removal prevent re-
459 splicing of mRNA. *PLoS Genet* **17**, e1009563, doi:10.1371/journal.pgen.1009563
460 (2021).
461 11 Wan, Y. *et al.* Dynamic imaging of nascent RNA reveals general principles of
462 transcription dynamics and stochastic splice site selection. *Cell* **184**, 2878-2895
463 e2820, doi:10.1016/j.cell.2021.04.012 (2021).
464 12 Johnson, B. S. *et al.* Biotin tagging of MeCP2 in mice reveals contextual insights
465 into the Rett syndrome transcriptome. *Nat Med* **23**, 1203-1214,
466 doi:10.1038/nm.4406 (2017).
467 13 Zhao, Y. T. *et al.* Long genes linked to autism spectrum disorders harbor broad
468 enhancer-like chromatin domains. *Genome Res* **28**, 933-942,
469 doi:10.1101/gr.233775.117 (2018).
470 14 Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse
471 tissues reveal evolutionary constraint in gene expression. *Nat Commun* **6**, 5903,
472 doi:10.1038/ncomms6903 (2015).
473 15 Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and
474 widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**,
475 1435-1440, doi:10.1038/nsmb.2143 (2011).
476 16 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with
477 applications to RNA splicing signals. *J Comput Biol* **11**, 377-394,
478 doi:10.1089/1066527041410418 (2004).
479 17 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence
480 logo generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).
481 18 Roy, M., Kim, N., Xing, Y. & Lee, C. The effect of intron length on exon creation
482 ratios during the evolution of mammalian genomes. *RNA* **14**, 2261-2273,
483 doi:10.1261/rna.1024908 (2008).
484 19 Blazquez, L. *et al.* Exon Junction Complex Shapes the Transcriptome by
485 Repressing Recursive Splicing. *Mol Cell* **72**, 496-509 e499,
486 doi:10.1016/j.molcel.2018.09.033 (2018).
487 20 Ferreira, M. A. *et al.* Collaborative genome-wide association analysis supports a
488 role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* **40**, 1056-1058,
489 doi:10.1038/ng.209 (2008).

21 Schulze, T. G. *et al.* Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. *Mol Psychiatry* **14**, 487-491, doi:10.1038/mp.2008.134 (2009).
22 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241, doi:10.1038/nature10945 (2012).
23 Iqbal, Z. *et al.* Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. *Hum Mol Genet* **22**, 1960-1970, doi:10.1093/hmg/ddt043 (2013).
24 Maruani, A. *et al.* 11q24.2-25 micro-rearrangements in autism spectrum disorders: Relation to brain structures. *Am J Med Genet A* **167A**, 3019-3030, doi:10.1002/ajmg.a.37345 (2015).
25 Brevik, E. J. *et al.* Genome-wide analyses of aggressiveness in attention-deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet* **171**, 733-747, doi:10.1002/ajmg.b.32434 (2016).
26 Shifman, S. *et al.* A whole genome association study of neuroticism using DNA pooling. *Mol Psychiatry* **13**, 302-312, doi:10.1038/sj.mp.4002048 (2008).
27 Lee, H. *et al.* Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am J Hum Genet* **90**, 746-751, doi:10.1016/j.ajhg.2012.03.004 (2012).
28 Sinha, V. *et al.* Variants in regulatory elements of PDE4D associate with major mental illness in the Finnish population. *Mol Psychiatry*, doi:10.1038/s41380-019-0429-x (2019).
29 Gurney, M. E. *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* **62**, 4884-4901, doi:10.1021/acs.jmedchem.9b00193 (2019).
30 Pan, T. *et al.* Dual functional cholinesterase and PDE4D inhibitors for the treatment of Alzheimer's disease: Design, synthesis and evaluation of tacrine-pyrazolo[3,4-b]pyridine hybrids. *Bioorg Med Chem Lett* **29**, 2150-2152, doi:10.1016/j.bmcl.2019.06.056 (2019).
31 Zhao, Y. T., Fasolino, M. & Zhou, Z. Locus- and cell type-specific epigenetic switching during cellular differentiation in mammals. *Front Biol (Beijing)* **11**, 311-322, doi:10.1007/s11515-016-1411-5 (2016).
32 Amid, C. *et al.* The European Nucleotide Archive in 2019. *Nucleic Acids Res* **48**, D70-D76, doi:10.1093/nar/gkz1063 (2020).
33 Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* **51**, 11 14 11-11 14 19, doi:10.1002/0471250953.bi1114s51 (2015).
34 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
35 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).
36 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
37 Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).

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

b

Loci	#	Loci	#	Loci	#	Loci	#	Loci	#
chr9_-_29674405	357	chr16_+_41206970	75	chr3_-_158361563	23	chrX_-_139990713	15	chrX_-_79478788	10
chr1_-_183339388	332	chr9_-_82916025	75	chr5_-_86060897	23	chr13_+_13755115	14	chr8_-_50162064	10
chr9_+_28029505	283	chr7_-_61241446	74	chr2_+_179627792	21	chr7_+_96232097	14	chr16_-_57444992	10
chr16_-_67364249	264	chr9_-_49719366	72	chr2_-_7282579	21	chr3_+_14131308	13	chr4_-_48248658	10
chrX_+_151836220	200	chr11_+_80386178	66	chrX_-_51473340	20	chr13_+_13753787	13		
chr11_-_77601762	197	chr16_+_40655810	64	chr18_-_12893721	19	chr14_+_32217381	12		
chr2_-_60076912	180	chr6_-_102503526	64	chr14_-_123590928	19	chr14_+_76119201	12		
chr14_+_119537648	171	chr9_+_47633924	60	chr7_+_87589088	19	chr1_+_180121414	12		
chr16_-_67142935	170	chr4_+_84984367	59	chrY_+_90808704	19	chr1_+_180121489	12		
chr15_-_73365196	146	chrX_+_112602663	54	chr2_-_39082928	19	chr19_+_47602798	12		
chr16_+_40262266	144	chr2_+_35124192	52	chr7_-_84005996	18	chr19_-_28354760	11		
chr10_+_69595048	143	chr1_-_156893013	45	chr4_-_56904127	18	chr12_+_102967315	11		
chr9_-_16029361	129	chr13_+_13755388	45	chr12_-_67059699	18	chr4_-_25728363	11		
chr9_-_49642259	113	chr13_+_13753651	44	chr3_+_159150264	18	chr8_-_99410596	11		
chr13_+_109333510	101	chr13_+_13754977	37	chr15_-_35088437	18	chr9_-_123653345	11		
chr12_-_104738849	91	chr13_+_109550007	33	chr2_-_136200636	17	chr1_-_123810502	10		
chr16_+_40979498	88	chr6_-_94050767	33	chr1_+_24684097	17	chr2_+_20490509	10		
chr16_-_74151684	83	chr1_-_55016183	30	chr13_+_13753821	17	chr14_-_73297836	10		
chr9_+_72309250	79	chr13_+_13755217	29	chr13_+_13753923	16	chr16_+_6158058	10		
chr6_-_37058429	77	chr13_+_109614417	28	chr6_-_148206211	16	chr12_+_65101518	10		

c

	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

d

Loci	#	Loci	#
chr7_+_123384756	9	chr9_+_113765922	8
chr2_-_125842760	9	chr14_+_105299009	8
chr9_+_76128920	9	chr6_-_108179985	8
chr1_-_154841063	9	chr13_+_96575819	7
chr11_-_37156217	9	chr1_-_97759169	7
chr10_+_21309092	9	chr1_+_11548381	7
chr2_+_177075009	9	chr1_-_183103753	7
chr1_+_7090857	9	chr9_-_112228936	7
chr11_-_11429127	9	chr11_-_47058958	7
chr11_-_77596357	9	chr1_-_136304718	7
chr13_+_13755081	9	chr2_+_102665261	7
chr8_+_20631694	9	chr4_+_24521096	7
chr1_-_13257060	9	chrX_-_85721958	7
chr8_+_88702858	9	chr2_+_20293447	7
chr8_-_128953205	8	chr3_+_40857854	7
chr2_-_37697912	8	chr2_-_64027968	7
chr14_+_105298953	8	chr11_-_76535247	7
chr11_+_77809797	8	chr12_+_82286608	7
chr5_+_108069164	8	chr8_-_50500205	7
chr7_-_98650773	8	chr2_-_39094563	7
chr16_-_56703624	8	chr14_+_14807267	6
chr13_+_13632543	8	chr14_+_122681487	6
chr17_+_21569014	8	chr15_+_22649279	6
chr11_-_33718266	8	chr4_-_36041544	6
chr6_-_43519253	8	chr7_-_34213648	6

Supplementary Figure S1: Novel RS sites. (**a**) The mapping statistics and access numbers of RNA-seq data utilized in this study. (**b**) The loci and numbers of junction reads (#) of the 84 RS site candidates. (**c**) 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. (**d**) The loci and numbers of junction reads (#) of the top 50 additional RS site candidates when lowering the cutoff of junction read count from 10 to 5.

Loci	#	Motif
5_+_109932925	245	AGGC
16_-_48972609	239	AGGC
6_+_104544118	201	AGGC
7_-_61276541	106	AGGC
5_-_107837447	90	AGGC
11_+_94024915	83	AGGC
7_-_61177531	53	AGGC
2_-_150343156	45	AGGC
6_-_37059005	43	AGGC
11_-_106869109	42	AGGC
6_-_135971965	42	AGGC
13_+_83876759	41	AGGC
16_-_31129626	37	AGGC
16_-_96059674	33	AGGC
16_-_4147165	31	AGGC
12_-_55695897	28	AGGC
2_+_158430383	26	AGGC
11_-_76440037	20	AGGC
18_-_34491113	20	AGGC
1_+_146564463	17	AGGC
1_+_105716092	17	AGGC
15_+_81617839	14	AGGC
6_-_128196834	13	AGGC
7_+_91367286	13	AGGC
7_-_62185448	13	AGGC
1_-_9079601	12	AGGC
11_-_81227425	11	AGGC
3_+_68114083	11	AGGC
6_-_135826847	11	AGGC
2_-_70784811	10	AGGC
1_+_128176822	781	AGGG
14_+_77365024	196	AGGG
3_-_56043120	132	AGGG
X_-_103588172	88	AGGG
6_+_140536606	84	AGGG
4_-_36122181	57	AGGG
4_+_109737141	44	AGGG
12_+_4970404	41	AGGG
6_+_105678014	41	AGGG
6_+_65085726	34	AGGG
8_+_14454399	34	AGGG
7_+_123188393	32	AGGG
9_-_95497294	26	AGGG
17_+_55463884	25	AGGG
9_-_104196331	23	AGGG
4_-_70383419	23	AGGG
2_-_18298980	22	AGGG
X_-_103605310	21	AGGG
2_+_28846761	20	AGGG
5_-_9702032	18	AGGG
14_-_7310298	17	AGGG
13_+_63017857	17	AGGG
13_-_93177259	17	AGGG
3_-_26246847	16	AGGG
14_-_119081379	16	AGGG
1_-_162075787	15	AGGG
14_-_103264399	15	AGGG
11_+_72869177	14	AGGG
16_+_5886118	14	AGGG
10_+_14086014	13	AGGG
9_+_66424938	13	AGGG
X_-_47843695	13	AGGG
14_+_60831447	11	AGGG
X_+_69407511	11	AGGG
2_+_147052659	11	AGGG
5_-_5644561	11	AGGG
11_-_26036909	11	AGGG
1_+_86812486	11	AGGG
9_-_106713666	10	AGGG
7_-_75255477	10	AGGG
8_+_79662433	10	AGGG
7_-_62100437	417	AGAT
4_+_102281769	142	AGAT
4_-_36409784	120	AGAT
12_+_4978451	118	AGAT
10_+_50635031	110	AGAT
15_+_35491126	97	AGAT
7_-_61195040	91	AGAT
18_-_22951206	89	AGAT
16_+_6636583	84	AGAT
13_+_91525613	82	AGAT

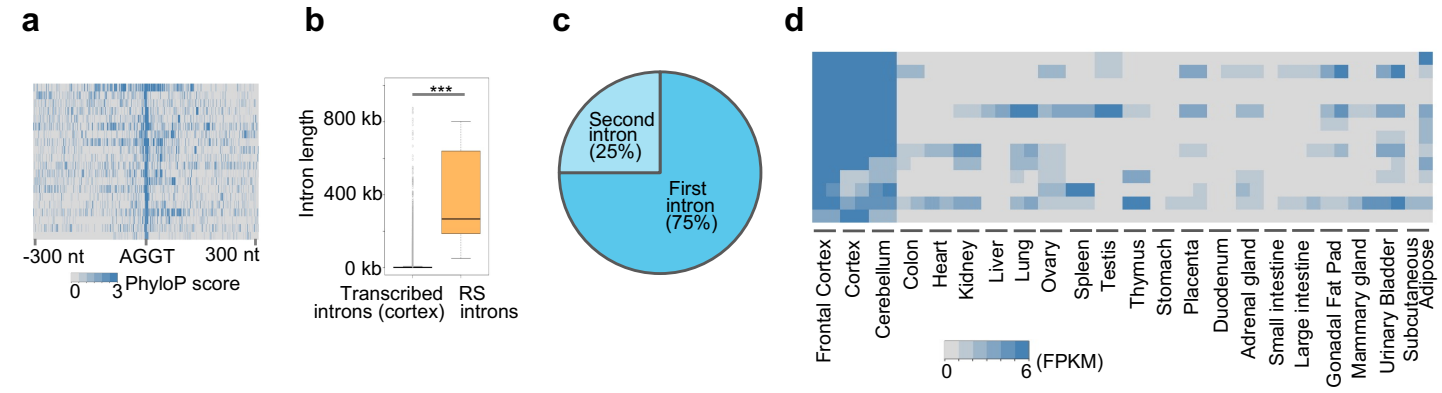
Loci	#	Motif
1_-_124753317	71	AGAT
9_-_10168834	63	AGAT
2_-_39150415	53	AGAT
4_-_126392561	38	AGAT
5_+_34842732	30	AGAT
3_+_159080765	29	AGAT
12_+_61604964	29	AGAT
12_+_3476947	29	AGAT
6_+_111239501	27	AGAT
2_-_146459719	26	AGAT
7_-_61233587	22	AGAT
7_+_63461661	21	AGAT
9_+_86479623	20	AGAT
17_+_55463017	20	AGAT
9_+_94873127	19	AGAT
2_-_113816737	17	AGAT
2_+_166049646	17	AGAT
1_+_115689399	17	AGAT
2_+_164081521	17	AGAT
7_-_84083045	16	AGAT
14_-_124574599	16	AGAT
1_-_66807605	15	AGAT
5_-_67781092	15	AGAT
X_+_155806465	15	AGAT
1_-_119779051	15	AGAT
6_-_102514241	15	AGAT
5_+_81105225	15	AGAT
6_-_37004587	15	AGAT
2_+_50369944	14	AGAT
13_+_104369652	12	AGAT
6_+_143192859	12	AGAT
5_+_150750195	12	AGAT
3_-_56045343	12	AGAT
16_-_17703150	12	AGAT
14_+_15006634	11	AGAT
11_-_37224567	11	AGAT
X_+_155927605	11	AGAT
X_-_111595382	10	AGAT
8_+_85896667	10	AGAT
2_+_18094951	10	AGAT
13_+_18808321	10	AGAT
2_-_162295753	10	AGAT
8_+_65721127	10	AGAT
12_+_88853103	761	AGAA
4_+_135858840	318	AGAA
3_+_35808607	106	AGAA
9_+_111374733	100	AGAA
12_+_52938338	96	AGAA
10_+_39742689	95	AGAA
4_-_35967938	89	AGAA
7_-_61243933	77	AGAA
1_-_8996619	76	AGAA
11_-_33404200	42	AGAA
7_-_110456965	37	AGAA
18_-_62521491	36	AGAA
11_-_36957961	30	AGAA
2_-_66411458	29	AGAA
10_+_29196224	29	AGAA
9_-_4518942	28	AGAA
2_+_147034828	27	AGAA
5_-_107843897	23	AGAA
16_-_45718909	23	AGAA
3_-_56016556	21	AGAA
4_-_126375477	21	AGAA
5_-_107811403	16	AGAA
10_+_115390236	16	AGAA
12_+_53640695	14	AGAA
2_+_48989026	13	AGAA
4_-_36365137	12	AGAA
4_-_151579857	11	AGAA
2_+_96911007	11	AGAA
13_+_91538752	11	AGAA
19_+_55340712	11	AGAA
9_-_29011413	10	AGAA
3_+_116012185	10	AGAA
16_-_60483911	97	AGAC
3_-_148861290	73	AGAC
9_+_110546662	54	AGAC
9_+_66418367	46	AGAC
13_+_108685992	29	AGAC
19_+_36982810	29	AGAC

Loci	#	Motif
15_+_35665136	27	AGAC
8_+_112803002	24	AGAC
3_+_40870094	21	AGAC
5_+_71760382	21	AGAC
15_+_39536794	16	AGAC
X_+_155736382	15	AGAC
15_+_35616038	15	AGAC
3_-_30909135	15	AGAC
4_+_83437585	15	AGAC
4_+_42919790	15	AGAC
8_+_66703006	14	AGAC
10_+_85008655	14	AGAC
16_-_37230076	13	AGAC
8_+_45739954	11	AGAC
X_-_19232047	11	AGAC
5_-_86153367	10	AGAC
15_-_71681548	10	AGAC
1_-_154799699	135	AGAG
7_+_87586513	127	AGAG
12_+_116350140	124	AGAG
9_+_121729995	123	AGAG
2_-_17525490	76	AGAG
7_-_61197854	61	AGAG
4_+_98469218	51	AGAG
X_-_102408966	49	AGAG
9_-_78569945	48	AGAG
12_+_116353825	48	AGAG
6_+_55406755	32	AGAG
2_-_66371099	29	AGAG
X_+_155843739	28	AGAG
11_-_109982269	28	AGAG
15_+_91768084	27	AGAG
9_+_92578417	27	AGAG
12_+_16935149	23	AGAG
9_+_113841157	22	AGAG
18_+_49921499	16	AGAG
9_-_10658288	15	AGAG
14_-_105140017	15	AGAG
2_-_34714295	13	AGAG
9_-_16017279	13	AGAG
4_-_87133582	11	AGAG
17_+_55490910	10	AGAG
15_-_48039542	10	AGAG
6_+_110761323	10	AGAG
9_+_111347464	342	AGCA
11_-_102537147	259	AGCA
17_+_11303588	151	AGCA
3_-_158263393	146	AGCA
5_-_25415027	109	AGCA
12_+_89191151	68	AGCA
2_-_23541616	61	AGCA
9_-_121257422	59	AGCA
5_-_106921077	37	AGCA
18_+_36626002	35	AGCA
6_+_81933726	31	AGCA
8_+_45688296	25	AGCA
4_+_86780291	25	AGCA
1_+_128260829	25	AGCA
8_-_48311696	24	AGCA
1_+_66593459	20	AGCA
4_-_76101589	17	AGCA
1_-_54181363	13	AGCA
12_+_116342043	13	AGCA
8_-_32441050	12	AGCA
11_-_33304011	11	AGCA
10_+_58906407	11	AGCA
8_+_35779842	11	AGCA
19_+_57673405	11	AGCA
8_-_17392675	10	AGCA
4_+_49638923	10	AGCA
2_-_17462857	113	AGCC
1_-_9249411	90	AGCC
11_-_77605639	77	AGCC
5_-_49385846	63	AGCC
7_-_61221933	61	AGCC
12_+_89103786	22	AGCC
14_-_119083959	22	AGCC
12_+_52783308	21	AGCC
18_-_61390087	17	AGCC
11_+_29210900	15	AGCC
X_-_102618379	12	AGCC

Loci	#	Motif
2_-_94328885	12	AGCC
9_-_107838931	12	AGCC
2_+_65668739	11	AGCC
8_-_48837059	11	AGCC
17_+_79618976	10	AGCC
1_+_66638582	46	AGCG
16_+_5953417	17	AGCG
X_-_103591907	181	AGCT
4_-_24693635	79	AGCT
8_+_90973922	70	AGCT
9_+_27336115	66	AGCT
11_+_80399427	35	AGCT
3_+_60616059	35	AGCT
2_-_6849356	33	AGCT
6_-_122483223	32	AGCT
11_+_23682506	27	AGCT
7_-_78325452	24	AGCT
1_-_25752435	20	AGCT
7_-_61151350	18	AGCT
5_+_16284880	15	AGCT
9_-_79679432	15	AGCT
11_-_65743466	14	AGCT
2_+_67843337	13	AGCT
4_+_150431889	13	AGCT
14_-_62733851	13	AGCT
3_+_156920055	13	AGCT
7_-_61279511	12	AGCT
9_-_53213140	10	AGCT
2_+_20407682	10	AGCT
4_+_21811465	10	AGCT
3_+_68064042	10	AGCT
3_-_85121870	10	AGCT
14_-_103220149	714	AGGA
1_-_52204517	517	AGGA
13_-_103057916	200	AGGA
15_+_22817626	142	AGGA
4_+_104280468	125	AGGA
2_-_59983971	120	AGGA
7_-_61279760	109	AGGA
16_-_37156845	106	AGGA
10_-_5412911	102	AGGA
16_-_96008215	93	AGGA
18_+_49903412	86	AGGA
6_-_114031474	58	AGGA
4_-_58872098	57	AGGA
2_-_18295709	52	AGGA
15_-_53657526	45	AGGA
9_+_86510613	45	AGGA
4_+_33323653	43	AGGA
2_-_66293896	41	AGGA
10_+_90069848	38	AGGA
3_-_120797291	33	AGGA
1_-_84809033	31	AGGA
7_-_61181703	30	AGGA
16_-_37225135	28	AGGA
15_-_24990196	26	AGGA
9_-_65000178	23	AGGA
7_-_61129946	21	AGGA
3_+_95451656	21	AGGA
2_-_62762877	19	AGGA
6_+_104730653	19	AGGA
1_+_63456102	17	AGGA
1_+_86820839	17	AGGA
1_+_159258645	17	AGGA
10_+_63249952	16	AGGA
19_+_29535172	16	AGGA
10_+_98926312	15	AGGA
2_+_97555336	14	AGGA
9_-_10532185	14	AGGA
14_-_23604448	14	AGGA
10_+_14101625	14	AGGA
4_+_104346080	12	AGGA
4_-_11295505	11	AGGA
1_-_25827430	11	AGGA
1_+_86813653	11	AGGA
16_-_16469373	11	AGGA
Y_-_1216512	10	AGGA
1_-_132542239	10	AGGA
2_+_102751345	10	AGGA
4_-_36434683	10	AGGA
14_-_12776073	10	AGGA

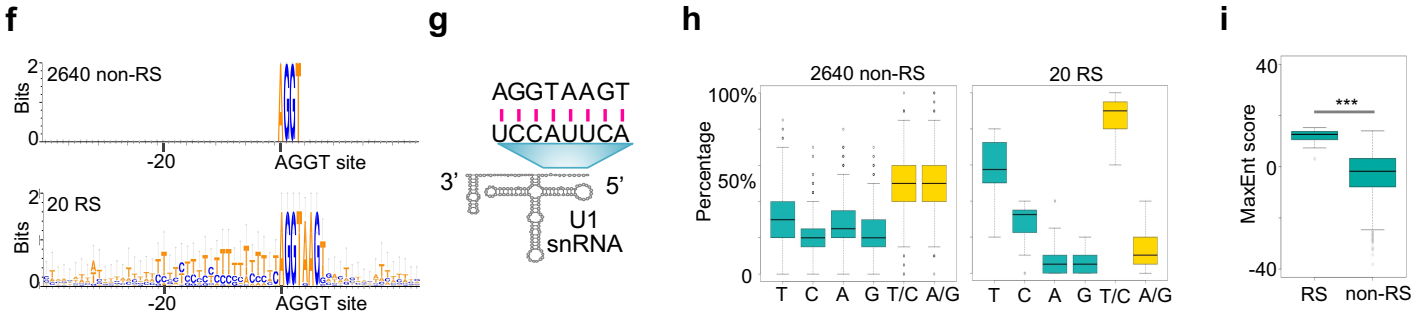
Loci	#	Motif
3_-_26132804	177	AGTA
2_-_112860532	139	AGTA
1_-_52217299	47	AGTA
12_-_98242533	39	AGTA
1_-_155961191	30	AGTA
5_-_128770054	19	AGTA
5_-_107814716	18	AGTA
15_+_8182520	17	AGTA
6_-_35067813	17	AGTA
6_+_45826164	16	AGTA
4_-_3562147	15	AGTA
10_-_49650033	11	AGTA
7_+_113553294	11	AGTA
1_+_86812510	116	AGTC
12_+_29708833	64	AGTC
4_-_36739098	46	AGTC
12_+_16934239	30	AGTC
10_+_106835175	27	AGTC
3_+_29311798	24	AGTC
14_+_27699639	15	AGTC
10_-_5175929	14	AGTC
16_+_43282969	14	AGTC
14_+_64731659	13	AGTC
2_+_20438936	11	AGTC
5_+_109927200	11	AGTC
2_+_102751260	10	AGTC
10_+_106828875	102	AGTG
2_-_39082853	45	AGTG
15_+_81415845	33	AGTG
15_-_38011943	21	AGTG
4_-_19539181	20	AGTG
6_+_110703440	18	AGTG
10_+_101866675	16	AGTG
8_+_65670877	14	AGTG
1_+_192226324	14	AGTG
9_-_119290866	13	AGTG
14_-_39470887	13	AGTG
2_+_37555478	13	AGTG
15_+_97789651	13	AGTG
9_+_120947066	12	AGTG
14_-_119063306	12	AGTG
6_+_106077480	11	AGTG
15_+_39394199	10	AGTG
4_+_110428203	10	AGTG
18_+_38454313	10	AGTG
4_-_36279317	237	AGTT
16_-_37241245	215	AGTT
10_-_89084654	58	AGTT
2_-_62845339	56	AGTT
12_+_61768006	52	AGTT
14_-_57643490	49	AGTT
16_+_6636524	43	AGTT
6_+_33307200	42	AGTT
7_-_6725524	40	AGTT
3_-_127244269	34	AGTT
4_-_76388148	33	AGTT
X_+_50884576	31	AGTT
4_-_36339631	28	AGTT
X_-_21059804	27	AGTT
1_-_123433116	25	AGTT
1_+_140383258	24	AGTT
11_-_29819185	23	AGTT
4_-_123479060	20	AGTT
13_-_59801259	20	AGTT
14_-_103292665	20	AGTT
6_+_119906384	20	AGTT
6_-_37016750	19	AGTT
5_-_36630577	18	AGTT
7_+_96969500	16	AGTT
2_-_136318423	15	AGTT
2_+_49073840	15	AGTT
5_-_45727743	15	AGTT
5_+_81173433	15	AGTT
9_+_72298680	14	AGTT
12_-_3308193	12	AGTT
15_-_35104421	12	AGTT
8_+_23694298	11	AGTT
9_-_10791688	11	AGTT
4_+_33392713	11	AGTT
14_-_12694475	10	AGTT
8_-_8444936	10	AGTT

Supplementary Figure S2: The 405 AGNN RS site candidates. #, number of junction reads.



e

	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

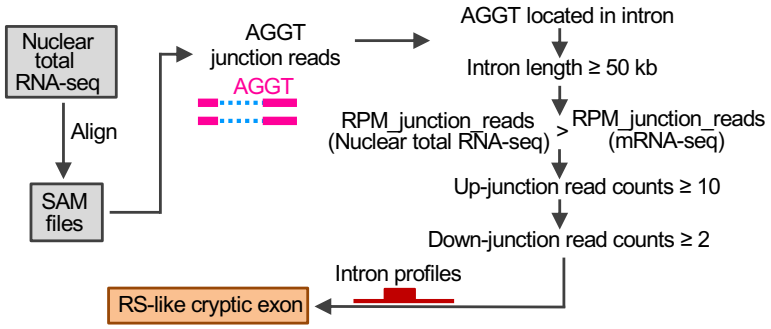


Supplementary Figure S3. Characteristics of RS sites. (**a**) Heatmap of phyloP score of RS sites and the flanking regions. (**b**) Boxplot of lengths of RS introns and introns transcribed in the mouse cortex. ***, $P < 0.0001$, one-tailed t-test. (**c**) Pie chart of locations of RS introns in host genes. (**d**) Heatmap of expression levels of RS genes in 22 mouse tissues. (**e**) The sequence motifs, nucleotide percentages, and 3'SS MaxEnt scores of the 20 RS sites. (**f**) Sequence logos of the 64 nt regions surrounding the 2640 non-RS AGGT sites and the 20 RS AGGT sites. (**g**) Schematic of the sequence base pairing between the AGGTAAGT motif and U1 snRNA. (**h**) Boxplots of the percentages of nucleotides in the 20 nt region upstream of the 2640 non-RS AGGT sites and the 20 RS AGGT sites. (**i**) Boxplot of MaxEnt 3' splice site (3'SS) scores of the 20 RS AGGT sites and the 2640 non-RS AGGT sites. ***, $P < 0.0001$, one-tailed t-test.

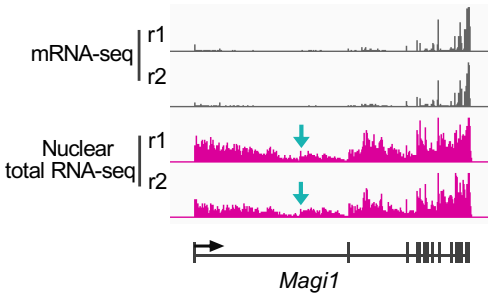
a

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

b



d



c

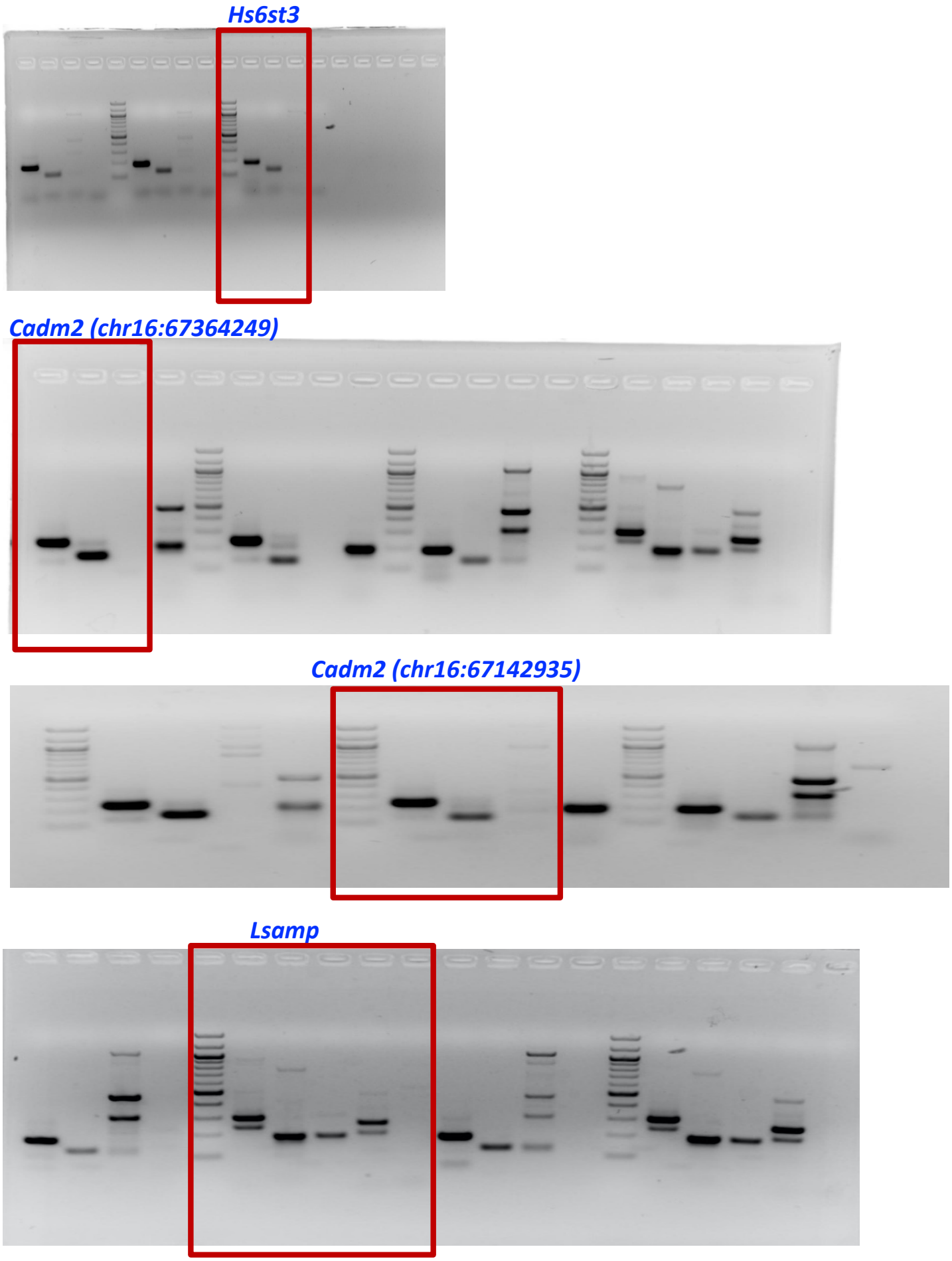
RS-like cryptic exon	Strand	Gene	Number of Junction reads		Upstream exon ending	r5'SS		Down 5'SS	
			Up	Down		Sequence	MaxEnt score	Sequence	MaxEnt score
chr2:17526050-17526130	-	Nebl	51	154	chr2:17529910	AATGTGGAA	-11.32	ATGGTAAGT	11.01
chr2:97522122-97522199	+	Lrrc4c	59	68	chr2:97468420	ACAGTGGAT	-7.51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

Supplementary Figure S4. RS exons and RS-like cryptic exons. (**a**) The genomic loci, numbers of junction reads, 5'SS sequences, and 5'SS MaxEnt scores of RS exons. (**b**) Schematic of the pipeline utilizing nuclear total RNA-seq data to identify RS-like cryptic exons. (**c**) The genomic loci, numbers of junction reads, 5'SS sequences, and 5'SS MaxEnt scores of RS-like cryptic exons. (**d**) Sequencing profile at *Magi1* locus. Green arrows indicate the putative RS AGGT loci.

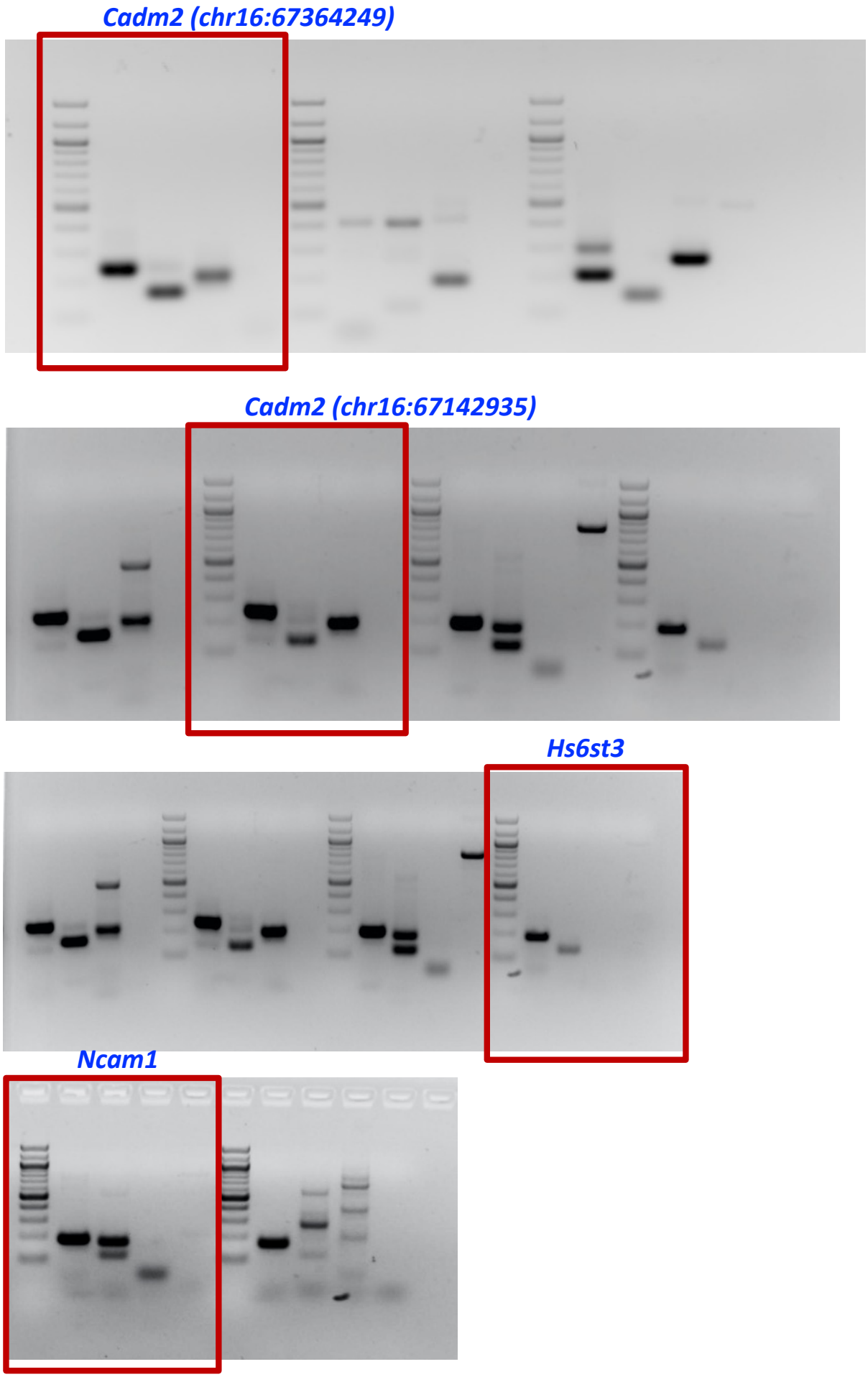
Primer name	Sequence (5' -> 3')
Cadm2 RS1 Forward Upstream-exon	GATCCTTCCCCAGACCCTTA
Cadm2 RS1 Reverse Downstream-exon	TGAGCTGGATTTGACCACTG
Cadm2 RS1 Forward RS-exon	GCTCCTTTATTAATGTCCTGTGG
Cadm2 RS1 Reverse RS-exon	CCACAGGACATTAATAAAGGAGC
Cadm2 RS1 Reverse Negative-control intron	CTATAAAAACAAGAAGGGGGA
Cadm2 RS2 Forward Upstream-exon	GATCCTTCCCCAGACCCTTA
Cadm2 RS2 Reverse Downstream-exon	TGAGCTGGATTTGACCACTG
Cadm2 RS2 Forward RS-exon	GTAAGCACATTAGTCATTCCTGT
Cadm2 RS2 Reverse RS-exon	TTGATGCAGTCTGATTATACAGG
Cadm2 RS2 Reverse Negative-control intron	CTAAAAATAAAAGGAAACAA
HS6ST3 Forward Upstream-exon	CTGGACCGAGCTCACCAAC
HS6ST3 Reverse Downstream-exon	CGTCGCACATATGGAGAGAGGT
HS6ST3 Forward RS-exon	CCTGTGTTGGAGTGACATACGA
HS6ST3 Reverse RS-exon	TCGTATGTCACTCCAACACAGG
HS6ST3 Reverse Negative-control intron	CTGAGATATGGGACAGAAGT
Ncam1 Forward Upstream-exon	GTCTGCTCCTCGGTCCA
Ncam1 Reverse Downstream-exon	CTCCTTGGCTGGGAACAATA
Ncam1 Forward RS-exon	GTAAGGAACTGCCAATCACG
Ncam1 Reverse RS-exon	CACGTGATTGGCAGTTCC
Ncam1 Reverse Negative-control intron	CTTTTGGTTGTGAGGTGAGAT
Lsamp Forward Upstream-exon	TGAGGGAAACCAAGGGAAGC
Lsamp Reverse Downstream-exon	CGCAAGCCCTACCTGAGGAT
Lsamp RS1 Reverse RS-exon	ACTCATCATTGCTGTGGGTGG
Lsamp RS2 Reverse RS-exon	TTTCTAAAGGACAGTCTGGGGA
Lsamp RS3 Reverse RS-exon	CTATGAGAATTGCCACCTGTGC
Lsamp Reverse Negative-control intron	TCTCTCCAGACTGATGGAGGAT

Primer name	Sequence (5' -> 3')
Lrr4c Forward Upstream-exon	TTGGTGTCAAGACCGTGCATTC
Lrr4c Reverse Downstream-exon	AGCCAACAGCACCACAAGC
Lrr4c Forward Cryptic-exon	CCATGGCCGGAATGGAATAAG
Lrr4c Reverse Cryptic-exon	CTTATTCCATTCCGGCCATGG
Lrr4c Reverse Negative-control intron	CTACAAATCCAAGAAGGGGA
Nova1 Forward Upstream-exon	AACTGGAGCCACTATCAAGCTG
Nova1 Reverse Downstream-exon	GATGCGATCAGGATTAACGGTG
Nova1 Forward Cryptic-exon	TGCCAGCGGCAGAAGAGAT
Nova1 Reverse Cryptic-exon	TGTGGTACTCCTAGTCAGTGTC
Nova1 Reverse Negative-control intron	CTAGGGGAAAGCGCGCTGTT
Magi1 Forward Upstream-exon	TCCAGAAGAAGAACCACTGGAC
Magi1 Reverse Downstream-exon	CATGGCGGTAGAGGTTGTC
Magi1 Forward Cryptic-exon	CAGATGCTGTCTTCCTGTGTTC
Magi1 Reverse Cryptic-exon	GTAGGCGAGCATATTGTGAACG
Magi1 Reverse Negative-control intron	CTGAAAAACATAAACACATA

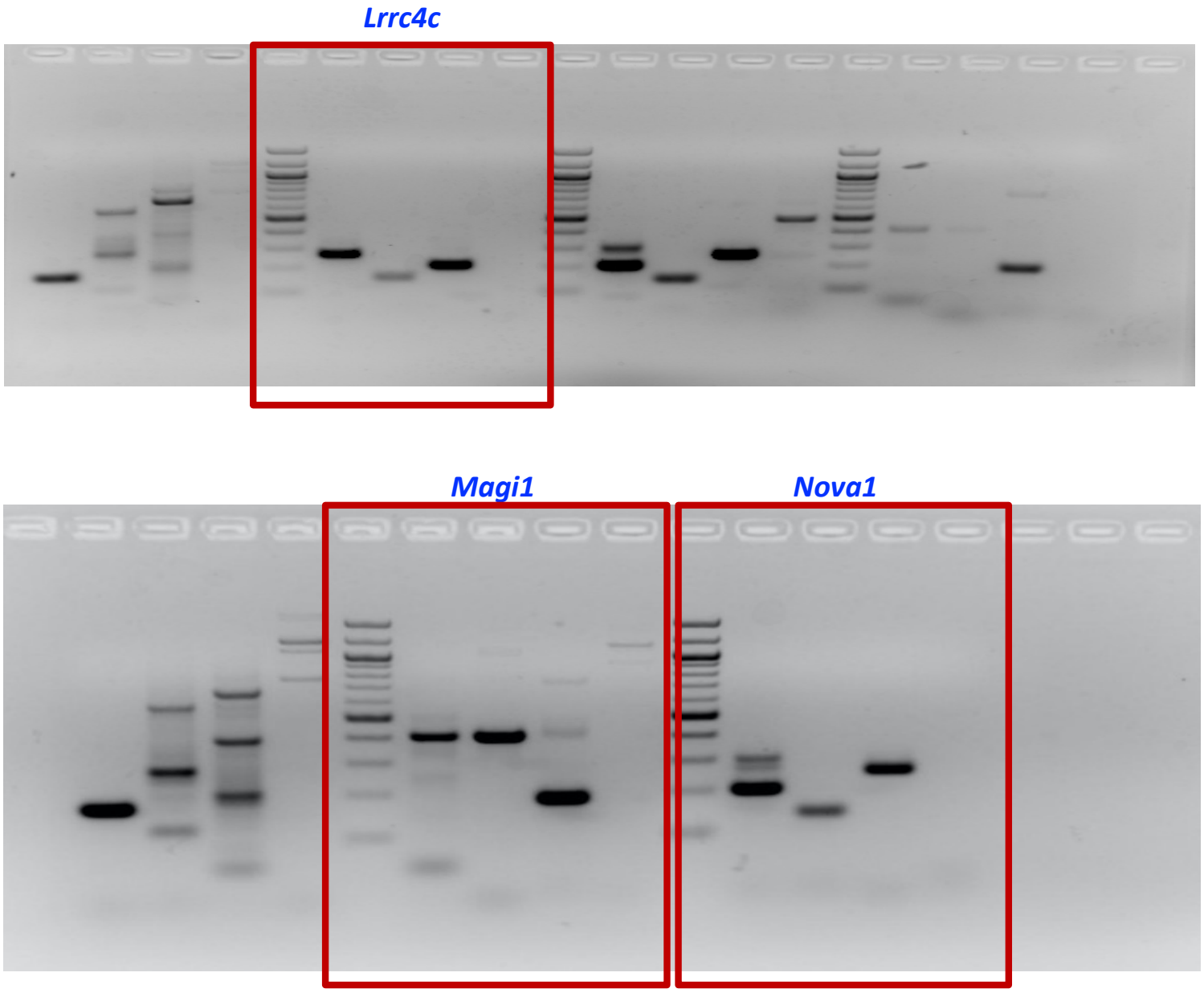
Supplementary Figure S5. RT-PCR primer sequences for the RS sites (top table) and RS-like cryptic exons (bottom table). *Cadm2* RS1: chr16:67364249. *Cadm2* RS2: chr16:67142935. *Lsamp* RS1: chr16:40262266. *Lsamp* RS2: chr16:40655810. *Lsamp* RS3: chr16:40979498.



Supplementary Figure S6. The full-length gels for plots in Figure 2.



Supplementary Figure S7. The full-length gels for plots in Figure 4.



Supplementary Figure S8. The full-length gels for plots in Figure 5.